DOCUMENT RESUME

ED 044 323	24	SO 000 277

AUTHOR	Sykes, Gresham M.; And Others
TITLE	Law and Social Science Research, A Collection of Annotated Readings. Volume II.
INSTITUTION	Denver Univ., Colo. Coll. of Law.
SPONS AGENCY	Office of Education (DHEW), Washington, D.C.
BUREAU NO	BR-6-1236
PUB DATE	69
GRANT	OEG-4-7-061236-0093
NOTE	472p.
EDRS PRICE	EDRS Price MF-$1.75 HC-$22.55
DESCRIPTORS	Curriculum Development, *Graduate Study, *Instructional Materials, Interdisciplinary Approach, *Law Instruction, Law Schools, Projects, Reading Materials, *Research Methodology, Social Sciences, *Sociology

ABSTRACT

A brief description of these materials is given in SO 000 276. This volume contains the last four sections of readings, comments, and class notes on: the problem of sampling, the problem of inference, measures of association, and data analysis. An 11 page bibliography of additional readings in methodology is appended. (SBE)

U.S. DEPARTMENT OF HEALTH, EDUCATION
& WELFARE
OFFICE OF EDUCATION
THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL OFFICE OF EDUCATION POSITION OR POLICY.

ED0 44323

LAW AND SOCIAL SCIENCE RESEARCH
A Collection of Annotated Readings

By

Gresham M. Sykes
Robert H. Sulnick
Norman K. Linton

COLLEGE OF LAW

UNIVERSITY OF DENVER

(Prepared with the assistance
of a grant from
The U. S. Office of Education
#OEG-4-7-061236-0093)

SO 000 277

TABLE OF CONTENTS

VOLUME II

CHAPTER FOUR: THE PROBLEM OF SAMPLING

Editors' Comments

Readings:

John H. Mueller and Karl F. Schuessler, *Statistical Reasoning in Sociology* (Boston: Houghton Mifflin Company, 1961), pp. 337-357.

Hans Zeisel, "The Uniqueness of Survey Research," *Cornell Law Quarterly*, 45 (1960), pp. 322-346.

Alfred F. Conard, *et al.*, *Automobile Accident Costs and Payments* (Ann Arbor: The University of Michigan Press, 1964), pp. 9-11; 322-366.

Paul Perry, "Election Survey Procedures of the Gallup Poll," *The Public Opinion Quarterly*, 24 (Fall, 1960), pp. 531-542.

Class Notes

CHAPTER FIVE: THE PROBLEM OF INFERENCE

Editors' Comments

Readings:

John H. Mueller and Karl F. Schuessler, *Statistical Reasoning in Sociology*, pp. 361-379; 388-391.

M. Finklestein, "The Application of Statistical Decision Theory to the Jury Discrimination Cases," *Harvard Law Review*, 80 (1966), pp. 338-376.

Class Notes

3

CHAPTER SIX: MEASURES OF ASSOCIATION

Editors' Comments

Readings:

John H. Mueller and Karl F. Schuessler, Statistical Reasoning in Sociology, pp. 238-249; 281-310.

H. L. A. Hart and A. M. Honore, Causation in the Law (Oxford: Oxford University Press, 1959), pp. 64-78.

Fowler V. Harper and Fleming James, Jr., The Law of Torts (Boston: Little, Brown and Company, 1956), pp. 1108-1121.

Morris R. Cohen and Ernest Nagel, Logic and Scientific Method (New York: Harcourt, Brace and World, 1934), pp. 245-272.

Class Notes

CHAPTER SEVEN: DATA ANALYSIS

Editors' Comments

Readings:

Hans Zeisel and Thomas Callahan, "Split Trials and Time Saving: A Statistical Analysis," Harvard Law Review, 76 (1963), pp. 1606-1625.

Allen H. Barton and Paul F. Lazarsfeld, "Some Functions of Qualitative Analysis in Social Research," Frankfurter Beitrage Zur Soziologie (Band 1, 1955), pp. 321-361. (Reprinted from Bobbs-Merrill Reprint Series in the Social Sciences, #S-336).

William J. Goode and Paul K. Hatt, Methods in Social Research (New York: McGraw-Hill, 1952), pp. 313-358.

Class Notes

APPENDIX A: Bibliography of Additional Readings in Methodology

4

CHAPTER IV

SAMPLING

I

In the great bulk of man's activities, one is forced to work with imperfect knowledge; we are so often ignorant of many of the key facts which would allow us to plan successfully and act with assurance. Our mastery of some aspects of the physical world leads us to forget how limited and how provisional our understanding of most of the world really is. And our statements about other human beings are often little more than wild guesses and pronouncements drawn from tradition.

One of the most common sources of man's difficulties in knowing about people, things, and events is that he is almost always dealing with a small fraction of the world that concerns him. He seldom assesses the whole--he usually must rely on a part, a sample. "In his continuing effort to adjust to the environment," note Professors Mueller and Schuessler in the first selection in this chapter, "man has necessarily been obliged to treat an isolated experience as typical of a larger system in order to profit by it."

The problem, of course, is how to select an isolated experience, a sample of people, things, or events, so that it best represents the universe from which it is drawn. Furthermore, there are the problems of measuring just how well we have performed this task and of making the best estimates we can about the sampled universe on the basis of the sample that is in our hands.

Statistical theory, based on mathematical laws of probability, provides some solutions to these problems. The use of random samples, purposive samples, stratified samples, cluster samples, and so on, as described by Mueller and Schuessler, are all techniques by which we can make statements about units in a population which we have never observed first hand--and attach a quantitative estimate of our accuracy.

I I

If ideas about probability and sampling are a commonplace in scientific inquiry (and, indeed in everyday human affairs, although often implicit or unrecognized), it is also true that they have not been readily accepted by the law. There are a number of cases, of course, in which data derived from samples have been presented in trial proceedings. And material drawn from the United States Census (which is heavily dependent on sampling procedures) has usually been allowed into the courts, as pointed out by Professor Zeisel in "The Uniqueness of Survey Evidence." But, in general, the law has been reluctant to rely on such facts and remains more than a little suspicious of the statistician and his talk of deviations, levels of significance, etc.

This suspicion on the part of the law may rest partially on a rather widespread antipathy toward quantitative modes of reasoning. Our legal system, after all, was developed at a time when few lawyers or judges or juries could have been expected to be very familiar with the mathematical concepts of statistics; and even today many people find the logic and the system of notation a serious barrier. It is also possible that lawyers and judges view theories of probability and the data associated with them as very suspicious objects, following the general line of thought that "there are lies, there are damn lies, and there are statistics." And then there is the possibility that the intellectual training of judges and lawyers directs their interest to the specific, the particular case, the unique event. Probability and sampling, however, involves classes of events, tendencies, the average behavior of groups of units rather than the actual behavior of any one unit; and the law may feel most uncomfortable when it is asked to move from the concrete reality of the individual to the abstract characteristics of a population as measured by a handful of examples.

In any event, whatever may be the general source of the law's intellectual and attitudinal reluctance to embrace the cold charms of statistics, there have been immediate difficulties in terms of the rules of evidence. We think that Professor Zeisel explores many of these issues. First, there is the problem of hearsay, since a good deal of the evidence involved is frequently drawn from surveys depending on verbal responses to questionnaires and interviews administered to large numbers of people. Second, there is the problem of a representative sample, since this sort of evidence is usually not based on the responses of every person in the population under examination. We think that Professor Zeisel is right when he says, "The law with respect to survey evidence is still far from settled doctrine. Thus far, the development has been guided by fears that, since most surveys are hearsay evidence, a bad one might too easily mislead the trier of facts. But . . . while the dangers of an uncritically received survey are real enough, they derive not from its hearsay character, but primarily from elements easily opened to expert review." Presumably, this expert review can be provided, just as lawyers and judges can become more aware of the relevance of data based on sampling. And when that happens the "orderly development of the law," placing a greater stress on such evidence, may well become a reality.

III

The problems of using data derived from samples are not confined to evidence for legal disputes, of course. Such data play an indispensable part in almost all social science research, since the limitations of time and money mean that we usually cannot examine the entire population which we wish to describe and analyse. And we must be just as careful in confronting the problems connected with hearsay and representativeness in research as in litigation.

The selection from Automobile Accident Costs and Payments is most important, at this point, as an illustration of how detailed and explicit the person who engages in research must be, if other people are to have adequate information to judge whether the sampling design is sound and whether the final sample is likely to be representative enough to justify generalizations. The authors of this study are careful to point out the problems connected with insurance company records and hospital records, in determining the universe under examination--all individuals injured or killed in automobile accidents that occurred in the state of Michigan during one calendar year. They note how police records also have their faults, but are probably the best source of data available when supplemented by court records. The authors tell us how many questionnaires in the sample were not completed and why. All of these issues and more are made painstakingly clear so that we have a basis for deciding if the sample can be considered a random one and if distorting errors have crept in as the study moves from a list of names on a piece of paper to a set of completed questionnaires and interviews. If we do not have this kind of information, we are left in the position of simply accepting or rejecting the author's conclusion on the basis of intuition, faith in authority, etc.

I V

The last selection in this chapter, "Election Survey Procedures of the Gallup Poll," by Paul Perry, is of interest because the problem it tackles is a very common one in designing social science research. Frequently, there is no master list of people, things, or events forming the universe to be investigated. Or, if such lists do exist, they are frequently out-of-date, seriously incomplete, or otherwise defective.

Selecting a random sample from an existing list or file of all the units in the universe being studied is relatively simple, since we can use a table of random numbers or take every nth unit.[1] When no such list or file is available, however, the problem of sampling grows more difficult and this is typically the case when we want to generalize to an entire community, to a segment of the population of the United States, or (as in the case of the Gallup Poll dealing with elections) to the voting population of the United States as a whole. The procedures used in such cases vary a good deal, but they often involve some form of the basic process underlying the procedure described by Paul Perry:

[I] See, for example, John G. Peatman, Introduction to Applied Statistics, (New York: Harper and Row, 1963), pp. 424-425.

☐ First, the division of the universe to be studied into a set of groups, geographical areas, etc., of smaller size.

☐ Second, the selection of a random sample of these smaller segments of the universe.

☐ Third, the selection of a random sample from each of the segments.

Essentially, all that is being done in this multi-stage sampling is to divide the universe into more manageable sections and then to proceed with the business of selecting a random sample--by first randomly selecting sections and then randomly selecting units within these smaller and more convenient subdivisions. The important thing is to maintain the basic idea of random sampling--namely, *that every unit in the universe has an equal probability of being selected*. If the number of units in each subdivision is not the same, we must take this fact into account.

The problem here is that if subdivisions containing *different* numbers of units are given the *same* probability of being selected, and if the *same* number of units is selected from each subdivision, we end up with a situation in which each unit no longer has an equal probability of appearing in the sample. (The chance of a unit appearing in the sample equals the probability of its subdivision being selected, multiplied by the probability that it will be selected from within the subdivision.)

There are two ways to solve this rather common difficulty: (1) We can give each subdivision an equal chance of being selected and then randomly select the same percentage of units from each subdivision; or (2) we can weight each subdivision by the number of units it contains, making the chances of a subdivision being selected proportionate to its size, and then randomly select the same number of units from each subdivision.

The main point, however, is simply this: Statistical data concerning social behavior is becoming of ever greater importance in all aspects of human activity, including making and applying the law. If the problems we encounter in this process are to be handled intelligently and justly, there must be an informed, cooperative effort among a variety of disciplines.

UNIVERSE AND SAMPLE I

The necessity of sampling in the social sciences is, we feel, made obvious by Mueller's and Schuessler's reference to a buyer sampling one peach from a bushel before making his purchase. It is not possible for him to test each piece of fruit and therefore he relies on a sample. The same is true for those who would inquire into human behavior. It is usually not possible to interview every student at a university, for example, so the researcher must also rely on a sample. There is, however, a basic difference between the kind of sampling which a buyer of peaches engages in, and the sampling procedures used in social science research. Scientific sampling is a body of technical procedures which is distinguishable by (1) its emphasis on a well-defined universe, and (2) a random selection of cases from that universe.

In statistical language the term universe refers to a totality of units possessing a well-defined common characteristic which determines membership in the set. Thus, such things as all American families in a given year or all students ever graduated from American colleges would constitute statistical universes. It is, however, the variable properties of the sampling units with which the scientist is primarily concerned. As Mueller and Schuessler put it: "We are seldom . . . interested in families qua families; rather our interest will lie in one or more of their relevant traits: income, nationality, size, social status, religion, and so on."

Defining a universe is the preliminary step in the sampling procedure. Then the sample must be drawn, and it is in the selection of sample respondents that we find a basic feature of social science research. There are two kinds of sampling procedures, random and non-random, and according to Mueller and Schuessler: "The ideal sampling procedure is one in which the drawings are affected by impartial chance factors alone, with the result that one item in the universe is as likely to be included in the sample as another." This, of course, is referred to as the random sample. Its advantage lies in the fact that no item is accorded preferential advantage of being selected, thereby avoiding all sorts of bias, the chances of selecting a representative sample are maximized, and it becomes possible to draw a number of statistical inferences.

Social scientists, however, often must compromise with the limitations of time and money and the nature of the research, all of which can make it virtually impossible to draw a true random sample. In effect what this amounts to is researchers making use of captive audiences such as boys on probation, selecting individuals whose last names begin with FA because "such names were probably randomly scattered in the various ethnic . . . groups," or taking every tenth name from the telephone book. These procedures have obvious utility for the researcher and are capable of providing worthwhile evidence, so long as the inferences drawn from them are accompanied by a full explanation of the relationship between universe and sample.

From John H. Mueller and Karl F. Schuessler, *Statistical Reasoning in Sociology* (Boston: Houghton Mifflin Company, 1961) pp. 337-357.

John H. Mueller and Karl F. Schuessler, Statistical Reasoning in Sociology (Boston: Houghton Mifflin Company, 1961), pp 337-357.

The Sampling Attitude

The Purpose of Sampling. The broad aim of statistics is to describe and summarize mass phenomena like births, deaths, and income, and their interrelationships. However, it is often necessary or practicable to base such description on a fraction of the total aggregate, and sometimes an exceedingly small one at that. Such an expedient may be, and usually is, quite satisfactory. It is astonishing how effective a well-selected fragment can be: a small snippet from a bolt of cloth; a few drops of blood from the patient's total supply; a few thousand survey votes, by which we describe the political intentions of millions of voters. Such procedure is standard practice in everyday social and economic life, as well as in the branches of scientific activity. In instances of this kind, when the data are partial rather than complete, and when they are used to characterize the entire set, we call the fragment a *sample*, and the total aggregate a *universe*, or *population*. We name a specified value of the universe, such as the mean, a *parameter*, and its counterpart in the sample we term a *statistic*. The objective of sampling is, therefore, to draw an inference about the parameter, which is unknown, from the sample statistic which is observed. This process of generalizing in a prescribed manner from sample to universe has come to be known as *statistical inference*.

Although statistical inference as a formal quantitative technique is principally an achievement of the twentieth century, its underlying motivation is as ancient as mankind itself. In his continuing effort to adjust to the environment, man has necessarily been obliged to treat an isolated experience as typical of a larger system in order to profit by it. Once stung by a bee, or burned on a stove, he is likely to view such objects as persisting hazards to be avoided. On the same principle, but in a more deliberate manner, he samples one peach from the bushel, plugs a watermelon, or gives the new car a trial run.

Moreover, fresh samples of experience continually stimulate us to reexamine our previous inferences. Thus, the success of one woman in Congress will cast doubt on the previously held opinion that women are politically incompetent; the grade records of Negroes in northern schools will force a revision of the stereotyped belief that Negroes possess an inferior mentality. Thus, we modify our premises in response to the unfolding sample evidence. In the language of statistics, we afford our initial hypotheses an opportunity to be revised or nullified.

While every man thus reasons informally from part to whole, statistical inference is much more rigorous than such mere folk practice. Sampling has come to be a body of technical procedures which must be deliberately applied and strictly adhered to if its goals are to be fully realized. Thus, the statistical universe must be well defined, and the sample must be properly drawn — two basic operations which are much more involved than appears on the surface. In fact, the sampling system, as it has evolved in the twentieth century, is essentially a technical accounting device to rationalize the procurement of information. By this criterion of efficiency, it is always preferred to complete enumeration when it yields data of requisite accuracy, since such a procedure is obviously more economical in a busy world of limited resources.

Further, the logic of sampling is reinforced by the generally accepted view of nature as being orderly and predictable. The discipline of statistics has not always enjoyed such a favorable climate of opinion, especially when applied to human behavior. In the mid-nineteenth century, Quételet and Buckle were denounced as materialists because they drew inferences from statistical records on the recurrences of crimes, suicides, and other human actions that ran counter to the then current doctrine of free will. It is therefore no accident that the sampling method has been cultivated most intensively and that the sampling mentality has flourished in the scientifically-minded, efficiency-oriented rational culture of the modern Western world.

Advantages of Sampling. The collection of sample data naturally requires less time and effort than does the compilation of complete data. Hence, surveyors of American public opinion universally avail themselves of samples of respondents, since results based on time-consuming total enumerations would be obsolete before they could be tabulated and published. Similarly, since 1940, sampling has been extensively employed by the U.S. Bureau of the Census in its decennial enumerations to provide more promptly detailed descriptions of various characteristics of the American population.

Since sampling is also less costly, it may be quite feasible when the financial burden of full coverage is prohibitive. Although a free-lance sociologist, for example, may believe it desirable to interview every divorced couple in a large city, he may find such a program impracticable to execute because of limited financial resources. He will then be content with a carefully selected sample. But this is not an unusual restriction. Research workers in general expect to arrive at valid generalizations on the basis of sample materials, presumed to be representative of some wider domain. Thus, the anthropologist interrogates a few native informants in order to reconstruct the entire pattern of a culture; the U.S. Children's Bureau relies on a sample of juvenile courts in order to establish national trends in delinquency.

Moreover, there are times when sampling is not only practically advantageous, but an almost inescapable necessity. When we destroy an object in the very act of measuring one of its characteristics, we must either sample only a few or destroy them all. This goes by the name of *destructive sampling*. For example, it is impossible to measure the life span of an electric light bulb without ending its life; consequently, to approximate the average life span of a large lot, the manufacturer necessarily resorts to a small sample. Analogously, the physician cannot afford to sample more than a few drops of blood without causing undue discomfort in his patient.

But destructive sampling may sometimes be abandoned, if we are unwilling to tolerate the destruction of even a single item. Where human life is jeopardized, we are not willing to risk the loss of a single sample case. And even when the sampling is not harmful, social considerations inhibit us from freely carrying out experimentation on human beings. It is only when an occasional life convict or conscientious objector voluntarily submits to a possibly harmful experimental treatment that public opinion permits destructive sampling to be carried out on human beings. In the usual case, such sampling is transferred to dogs and monkeys, who serve as stand-ins for man.

Sample materials are thus the stuff of which scientific generalizations are made, but the sampling method involves more than mere recognition of that principle. Modern sampling practice is distinguished by (1) its emphasis on the well-defined universe, (2) the random selection of cases, and (3) the estimate of the reliability of the sample statistic — that is, how closely it probably conforms to the unknown parameter. These concepts form the basis of the discussions in this and the following chapter.

The Statistical Universe (Population)

Definition of Universe. In colloquial speech, the term *universe* suggests the entire Creation. But in statistical language, it refers merely to a totality of values possessed by elements having a well-defined common characteristic which determines membership in the set. Such sets are also termed statistical *populations*. We may cite as universes: the incomes of all American families in 1960, the opinions of all college students on the subject of war, or the social-status ratings of all residents in Yankee City. From these examples, we discern that the statistical universe may be conveniently conceived of as being dualistic in nature: (a) one dimension consisting of the units (e.g., families) which are actually sampled and which are therefore called *sampling units;* (b) the other dimension being the *sampling trait* (e.g., income) possessed by the sampling units, which is subsequently manipulated statistically.

It is the sampling units that are physically selected: families with incomes, farms with acres, students with opinions, workers with occupations. Further, the well-defined criterion which determines eligibility for such selection here applies to the sampling units rather than to the traits. Thus, a population of families consists of all human groups that satisfy the working definition of a family.

On the other hand, it is the variable properties of the sampling units that command our ultimate interest. We are seldom if ever interested in families *qua* families; rather our interest will lie in one or more of their relevant traits: income, nationality, size, social status, religion, and so on. It is variables such as these that are subjected to statistical measurement after the sampling units have been drawn and their characteristics determined.

When, for any reason, it is unnecessary to identify explicitly both aspects of the universe, we may quite properly employ elliptical statements such as "the population of college students" or "the universe of attitudes." But such abbreviated statements omit the sampling trait of attitude (of the college students) in the first instance, and the sampling unit, the college student (whose attitudes are polled), in the second instance. Although terminology varies among statisticians, we here conceive of the "universe" as comprising both the sampling unit and the corresponding sampling trait.

Universe Classification. Statistical universes differ in various ways and may therefore be classified according to various criteria. Three of the more important classifications are here briefly set forth to amplify the sampling concept and refine its implementation:

(1) The universe will be *qualitative* or *quantitative* according to whether the traits of which it is composed are attributes or variates. This familiar distinction implies that the statistical description of the universe will take the form either of arithmetic averages or frequency counts and percentages. For example, the incomes of American families may be represented by the mean, whereas the occupations of American workers can only be grouped and counted and expressed as percentages of the total.

(2) The universe may be *finite* or *infinite*, depending on whether the sampling units are finite or infinite in supply. By definition, a finite uni-

verse contains a countable number of elements. It may be relatively small, as, for example, all students enrolled in a particular college in a given year; or it may be relatively large, as all college enrollees in the United States. But the infinite universe consists of an endless number of elements, such as an unlimited number of penny tosses or other experimental trials. It is thus purely conceptual, and may even seem metaphysical to the finite mind. And yet it is often heuristically postulated in statistical inference. For one reason, an infinite population permits the reliability of the sample findings to be more simply evaluated by formula than does a finite population. Consequently, we resort to the assumption of an infinite population whenever the size of a finite population is large enough to justify it. Additionally, it may be invoked when it is not reasonable to limit the size of the universe at all, as in the case of infinitely repeatable experimental trials. The laboratory dog may die after an experimental injection; but the scientist does not restrict his generalization to the dead dog. His ultimate interest lies in the potentially endless succession of experimental trials ideally performed under identical conditions. Onto this infinitely large universe he fastens his generalization. Analogously, but at considerably greater risk, the social scientist may conceptually extend his findings in one or a few cases to the hypothetical infinite class of presumably identical events. One family, one community, one culture, one bureaucracy — each in its turn serves as a prototype to which all future occurrences of the same general class are presumed to conform.

(3) The universe which is actually sampled (the *sampled universe*) will not always coincide with the universe on which our sights are fixed (the *target universe*). The target universe represents ideally the territory we intend our generalizations to cover — the domain to which we eventually apply our sampling knowledge. Our ultimate interest may lie, for instance, in the patterns of adjustment of all married students on campuses in the United States, but for practical reasons it may be necessary to restrict the sampling to the available couples on a particular campus, and these couples then become the sampled universe. When a mailed questionnaire is returned by less than 100 per cent of the sample, as is usually the case, we may conceive of the target universe as the complete mailing list from which the sample was selected, and the sampled universe as all persons on that list who would theoretically return that questionnaire if given an opportunity to do so. Thus, the concept *target universe* may be applied broadly to an idealized extension of the sampled universe, or narrowly to a universe in which a fraction of the units are for one reason or another inaccessible to measurement, such as refusals or not-at-homes.

Strictly speaking, statistical inference should be rigorously limited to the sampled universe; and yet the social analyst can hardly refrain from speculating about his target universe. No research study in the social sciences would ever be made if its findings could not be imaginatively extrapolated beyond the limited universe from which the sample has been derived. Such speculations are both justifiable and desirable, provided that their tentative nature is recognized and understood. A sampling study of mental health and social class in a New England community is of significance primarily for the light that it sheds on the relation between social structure and personality deterioration in other American communities. And so the investigator quite naturally probes his sample materials for their wider generality and makes the most of his costly data. In projecting his findings onto the vaguely defined target universe he necessarily proceeds without benefit of strict reliability procedures. Yet he may be, and usually is, engaged in fruitful and necessary scientific activity. Nevertheless such a liberal statistical morality, which is practiced by even the most thoughtful and productive social scientists, is by no means license for irresponsible and sloppy statistical generalizations. It should be permitted only to experienced investigators.

Sampling Procedures

Problems of Sampling. The process of sampling is in its scientific sense a technical operation which must be conducted according to standard prescriptions in order to secure all of its benefits. In fact, costly social investigations have sometimes been severely blemished because the sampling tactics were crude and inadequate. Because of the admitted difficulties in sampling human populations, the discussion of the theory and practice of sampling must always occupy an important place in the domain of social statistics.

Consider, by way of illustration, the task of sampling a set of college students — a common assignment for majors in journalism or sociology — with a view to generalizing about the entire student body. Disregarding momentarily the kind of data sought — attitude toward communism, number of dates per week — how may we obtain a sample which will do justice to the student body? This is no simple task. The sampler might give way to the first impulse to take all persons enrolled in elementary sociology — a rather attractive possibility since such classes include a wide variety of students and, in addition, are easily reached and manipulated for experimental or survey purposes. However, on second thought, he will realize that college courses are almost sure to exert some selective influence among students. Thus, sociology is likely to attract persons whose primary interest is in social issues, but may hold no appeal for those principally interested in the physical world. In restricting the sampling to sociology students, there is therefore a danger of excluding certain types of individuals who are negatively selected by this subject. To forestall this outcome, we might propose using persons enrolled in English Composition, as this subject is required of all. Again, reasonable enough on first thought, this alternative will seem less satisfactory when we realize that only underclassmen are enrolled; mature upperclassmen would have no opportunity to be included in the sample.

Many other possibilities will readily suggest themselves in the search for representative coverage. Thus, we might consider canvassing the men's dormitories. But this scheme rather glaringly omits women, and so would have to be modified to include women's residence halls as well. Even with this modification, the plan is rather obviously unsound: it makes no provision for students housed in fraternities or for those living at home. Apparently, other plans will have to be tried until an adequate one is developed.

The general shortcoming of the aforementioned alternatives will be recognized by even the casual reader: they do not afford each sampling unit an equal opportunity of being selected. Ideally, of course, our sampling procedure should give to each unit in the population such an opportunity. Methods which meet that criterion are named *random sampling procedures;* methods which offer no such assurance may be conveniently labeled *non-random procedures.* When such equal opportunity of selection is not provided for by the sampling technique, it will generally be impossible to vouch for the representativeness (reliability) of the sample, and consequently it will be impossible to generalize confidently from the sample back to the universe. The fulfillment of the purpose of sampling is jeopardized.

Non-random procedures are therefore seldom considered to be ideal; nevertheless, they are often justifiably resorted to in social research because of practical necessity. Beginning with the least useful, several are presented here in ascending order of utility so that the student will be familiar with their respective merits and shortcomings.

Non-random Procedures

Haphazard Sampling. The acceptance of whatever cases one fortuitously happens to encounter, without any consideration whatsoever for their degree of representativeness, may be termed *haphazard* sampling.* This practice is exemplified by the old-fashioned straw vote in which citizens are accosted in the street to ascertain their voting intentions. From these straws in the wind, the election "forecast" is made. The obvious objection to such casual procedure is that the man-in-the-street simply is not representative of the total electorate. Similarly, in a survey of student opinion on the propriety of final examinations, those nighthawks whom we happen to find breakfasting in the coffee shop at 11 A.M. will be something less than a fair cross-section of the entire student body. Such crude chunks of data are "samples" only in the loosest sense of that term. When seriously used, they constitute an unflattering reflection on the sophistication of those who resort to, and accept, such data.

---

* Synonymous concepts that appear in statistical writing include *accidental sampling* and *convenience sampling.*

Availability Sampling. Although fully aware of the limitations of nonrandom sampling, the experienced social scientist will sooner or later realize that some form of it is often the only alternative to abandoning the inquiry. In many instances, he may be required therefore to seize whatever opportunity is available. The Kinsey survey of sexual behavior in the United States male population was severely criticized because it was based to a large extent on data provided by solicited subjects who made themselves available. It was plausibly contended that persons who volunteer to provide information about their sexual behavior are likely to differ significantly from persons who decline to be interviewed. But the investigators considered the acceptance of volunteer respondents as a pragmatic solution, rather than the preferred procedure. It was their reasonable assumption that many random selectees could not be induced to relate their sex histories, and that any possibility of obtaining a completely random sample was precluded.

For similar reasons, other types of sociological studies must often rely on available opportunities. For example, it may be impossible for the behavioral scientist to sample all workers in a given industry, but yet possible to observe workers in a local plant; impossible to poll a group selected from all school children in the state, but possible to poll those attending various local schools. Social psychologists are likely to find little humor in the sarcastic remark that their broad "universal" generalizations are founded on experiments on college sophomores. Above all, they would prefer to study all kinds of individuals — old as well as young, non-college as well as college — in formulating the laws of social learning. But it is often a question of sampling the available students or otherwise abandoning the project. Hence, they must convert a captive audience of students into a "valid" sample. Similarly, research on delinquency is almost always based on the available delinquents — boys on probation, boys in court, or boys in the industrial school. Regardless of how desirable it might be to study a group selected from all delinquents, the unapprehended as well as the apprehended, such comprehensive sampling is for obvious reasons out of the question.

It would therefore be pedantic to deny the uses of available opportunities, even though they do not yield ideal data. Social scientists, like everyone else, must often content themselves with compromises. Notwithstanding their shortcomings, availability samples do yield significant information and insights. For example, the Burgess and Cottrell sample of 526 married couples produced many revealing propositions on the positive and negative factors in marital happiness; Znaniecki's renowned sample of letters written by Polish peasants provided suggestive hypotheses on basic human motives. Moreover, inexpensive captive samples — college classes, prison inmates, members of the armed forces — can often be utilized to develop research techniques and to supply experience in applying them. No one would seriously protest against studying hospital patients because they do not perfectly represent all diseased persons; few would abandon the study of prison inmates merely because they are not an exact replica of the criminal population. Availability samples are quite legitimate, so long as the inferences drawn from them are accompanied by reservations which are made necessary by the ill-defined relation between universe and sample.

Judgment (Purposive) Sampling. When the composition of the universe is known to the sampler, and when a small sample of only one or a few items is required, the sampler may elect to rely on his mature, sound judgment in choosing the sample. Judgment sampling will therefore be most effective in the hands of an expert who knows his population, and who can readily spot the typical case. Why should he choose by lot, and thereby run the risk of drawing an extreme case, when he is confident that it is possible to obtain a representative item by deliberate choice? He legitimately pits his personal judgment against the operation of the laws of probability. The Lynds did not leave the choice of Middletown to blind chance; for, if they had done so, a highly atypical community might have presented itself for study and defeated the very purpose of the investigation. These sociologists were interested in a single modal community, as indicated by their choice of pseudonym. Judgment sampling is therefore especially well-suited to the *case study*, in which many aspects of a single representative case are observed and analyzed.

While judgment sampling is considered efficient for selecting only one or two items, it has also been applied on occasion to procure samples of considerable bulk. The *Time Magazine* sample of U.S. college graduates (1952), consisting of individuals whose last names began with "Fa," was justified by the surveyors on the ground that such names were probably randomly scattered in the various ethnic, religious, and socio-economic groups of the population. And since a genuine randomized selection would have required a much more cumbersome sampling apparatus, along with trained hands to operate it, the "Fa" type of sample was more practicable and hence in this instance the preferred alternative.

However, the seductive plausibility of judgment sampling should not be permitted to conceal its hazards. Even the expert's judgment is subject to human bias and error. And even when the universe is in plain sight, it is probable that the observer will misjudge to some extent the representativeness of a sample. To demonstrate such bias, the English statistician Yates once requested 12 persons to select three samples of 20 items each from a collection of 1,200 stones, each sample to represent as accurately as possible the size distribution of that experimental universe. Although the observers were free to view the collection at their leisure, still there was a consistent tendency to exaggerate the average size of the stones and to minimize their variation. In short, there was a constant error in judgment. We cannot be sure that analogous errors did not bias the "Fa" sample. Because of the practical certainty of human bias, judgment sampling must therefore be applied with great caution. But whatever the limitations of non-random samples may be, the techniques of descriptive statistics must obviously be competently applied in order to assure their maximum utility.

Random Sampling Procedures

Definition of Random Sampling. If available samples are fallible, and expert judgment is not to be trusted, then what are the factors that should determine the composition of the sample? The answer is: chance factors. However, in view of the low regard in which chance factors are usually held as guides to action, it is surprising that we should so willingly lay aside our cumulated knowledge and experience and go to the other extreme, permitting blind chance to determine the choice of the sample. In most human situations, we wish to eliminate chance, since it disturbs our predictions. Nevertheless, the ideal sampling procedure is one in which the drawings are affected by impartial chance factors alone, with the result that one item in the universe is as likely to be included in the sample as another. No item is accorded a preferential advantage. In fact, *random sampling* is defined as a procedure that provides an equal opportunity of selection to each unit in the population. We eliminate personal bias by elimination of the person, and thereby permit only the play of impersonal chance forces.

There is nothing esoteric about random sampling; most persons are familiar with a few homemade routines that assure an impartial randomized selection. The thoughtful hostess wishes every guest to have an equal opportunity to win the door prize, regardless of wealth or friendship; the contest is to be perfectly democratic. Accordingly, everyone is requested to write his or her name on a blank slip and drop it into the hat. After the last arrival, the slips are thoroughly mixed, and one name is drawn and declared the winner. Although only one person can win the door prize, we still say that the probability of winning was identical for all. Among 25 guests, the probability of winning would be 1 in 25. How is this to be interpreted? The layman would reply: "Since the human sampler cannot influence the drawing, each guest has the same chance." But the statistical interpretation lies in the very definition of probability, which applies to mass phenomena or conceptually repeatable trials. In this instance, "same probability" refers to the expectation that each name would be drawn an equal number of times if identical drawings were conducted indefinitely. Of course, an empirical demonstration of innumerable drawings is not undertaken; nevertheless, it is tacitly assumed that the a priori expectation of equal occurrences would eventually be confirmed. This principle is intuitively recognized by the losing guests who confidently console themselves with the bromide, "better luck next time."

Of course, equal probability of selection cannot be assumed if some cases are less accessible than others or if the selection mechanism is functioning imperfectly. Either of these related contingencies will defeat the aim of random sampling and result in *biased sampling*. Thus, in selecting a sample of 10 slips from a receptacle containing 200, randomness would be precluded if the slips were unequal in size, shape, or weight; or if they were carelessly mixed so that the last names dropped into the hat would have the best opportunity of being drawn. But the outward appearance of the extracted sample would never reveal such procedural flaws, or biasing factors. The composition of a non-random sample is visibly no different from that of a sample of the same size selected by random procedures; the sampling operation leaves no telltale mark. Thus, if we were to come upon two different scatters of ten coins each strewn on two tables, the one all heads and the other showing six heads and four tails, no amount of visual inspection would tell us whether either or both of them had been carefully laid down or whether they had been tossed at random. Nevertheless, the observer will intuitively conclude that the first had probably been laid down and the second had been arbitrarily tossed. Why? Because the probability of 10 heads together is so small, and the division of 6 and 4 much greater. In fact, the probabilities compare as 1 to 210. Still, he can never be certain. Similarly, in the case of our hostess, the name on the winning slip can never betray whether the drawing was honest or biased.

However, a suspicion of biased sampling will be aroused by the subsequent discovery of a marked discrepancy between the sample value and the true value of the universe, when and if the latter becomes known. The now classic example of such a melancholy outcome is the notorious *Literary Digest* Presidential pre-election poll of 1936. While Roosevelt obtained approximately 60 per cent of the popular vote in the actual election, his percentage in the *Literary Digest* sample of over 2,000,000 respondents was only 40 per cent — a difference of 20 percentage points! This discrepancy in such a large sample was symptomatic of a gross defect in the sampling procedure, which, upon later review and analysis, statisticians easily established. Although such a discrepancy could have occurred by chance, it would have been extremely unlikely. Hence the deduction that the method of sampling was biased. But such appraisals are always retrospective. In advance of the sampling, one can only provide for adequate machinery which will be reasonably certain to yield a random selection.

Simple Random Sampling. If called upon to devise a do-it-yourself sampling technique, the inexperienced layman, like our hostess, would probably procure as many slips as there are items in the population to be sampled. Slips of paper are more easily shuffled than people. We may surmise that he would then number these slips consecutively from 1 through N, corresponding to the numbered units of the universe. Next, he would place the slips in a suitable receptacle and mix them until the set was thoroughly scrambled. Finally, he would reach in and take out as many cases as desired. If he wished a sample of 10 cases, he would take out 10 different slips. Statisticians recognize this as the simplest type of random sampling and have therefore dubbed it *simple random sampling.*

They do not casually describe it as "reaching in and drawing out n different items." Rather they define it as that procedure in which every distinct sample of n items has an equal probability of selection from a finite population of N items. This definition expresses the long run consequence of "reaching in and taking out n different items." For, if that procedure were applied indefinitely to a given population (restoring the entire sample after each trial), each different sample would tend to reappear an equal number of times. In its procedural aspect, simple random sampling is the apparatus that guarantees the fulfillment of this criterion of equal probability. In its substantive aspect, it is the very criterion itself.

To unfold further the meaning of simple random sampling, let us consider the number of ways in which samples of 2 items can be selected from a miniature population of 5, whose members we shall designate: a, b, c, d, e. By manipulation, we discover that there are 10 different possible combinations, or samples:

ab	bc	cd	de
ac	bd	ce	
ad	be		
ae			

Hence, the probability of each combination in simple random sampling must be 1 in 10; each sample is expected to occur on the average once in every 10 trials.

For finite universes and samples of any size, this probability may be expressed in terms of the now familiar combinatorial formula:

$$Pr \text{ (Given Sample)} = \frac{1}{C_n^N}$$

Applying these formulas to the illustrative data given above, we have

$$C_n^N \text{ (Total Possible Samples)} = \frac{5 \times 4}{2 \times 1} = 10$$

$$Pr \text{ (Given Sample)} = \frac{1}{10}$$

which agrees with the previous results.

Sampling by Random Digits. A rudimentary technique for carrying out simple random sampling has already been set forth: (1) represent units on slips; (2) thoroughly scramble; and (3) draw the required number of slips. However, the drawing need not be laboriously carried out in this manual fashion. It will usually be inefficient to do so, especially when the population is large. It is much more practicable to substitute a table of randomly ordered digits for the shuffled numbered slips — a common procedure of research workers. This technique necessarily requires that we number serially all of the units in the population from 1 through N and then draw from the corresponding table of random digits as many different numbers (combinations of digits) as there are cases to be included in the sample. The cases whose serial numbers correspond to those drawn from the table constitute the sample. Table IV of the Appendix presents such a list of random numbers.

Other Random Sampling Procedures

Simple random sampling is the most primitive and unrestricted selection procedure and most clearly exposes the essential operation of randomness. Being free of procedural modifications which are often made necessary by practical circumstances, it is conceptually the simplest of all sampling routines and is therefore so labeled. Since it is random sampling in its most uncomplicated form, it serves as a standard of sampling efficiency against which other types are compared and evaluated. These other types are made necessary by the fact that simple random sampling in its pure form can almost never be employed in large-scale social research. It is far too impractical and costly, and often even impossible. Nevertheless, it is essential that the student clearly comprehend its basic characteristics in order to be able to recognize and appreciate the degree to which the alternative methods depart from this standard model. Three of the most prevalent alternative types are here set forth as a brief introduction to the subject: (1) *stratified,* (2) *cluster,* and (3) *interval* sampling. These sampling devices themselves are largely an outgrowth of the sheer practical problems that have arisen in the sample surveys of large human aggregates, which accounts for their wide currency in social science.

Stratified Sampling. "Do the citizens of Brownville favor racial integration of schools?" No competent surveyor of public opinion would attempt to answer that question without canvassing both white and Negro residents of the community. Public sentiment on this issue would not be reliably portrayed by a sample that slighted either group, since their opinions are so divergent. Yet such an imbalance between groups might occur under simple random sampling, unless the sample were made large enough to forestall that eventuality. A more economical alternative would be to sample each subgroup separately and combine the results, and thereby avoid a costly inflation of sample size. Such an operation, which first separates the entire population into relevant *strata* before randomly drawing the sample, is known as *stratified sampling.*

From a procedural standpoint, stratified sampling therefore consists of the following stages: (1) division of the total universe into subclasses, or *strata;* (2) the selection of a random sample from each stratum; and (3) the consolidation of the subsample statistics into a combined statistic weighted for size of strata. In this context, the term *stratification* does not, of course, connote a hierarchy, as the ranks of an army, or the geological seams of the earth crust; rather it signifies the categories of a statistical variable, such as race, sex, or religion, into which the total population is conveniently divided.

Nor does stratification imply a relaxation of the requirement of randomized selection, although that inference has been sometimes mistakenly drawn. This misconception may possibly reflect a failure to distinguish clearly between stratified random sampling and so-called *quota sampling,* which was formerly widely used in opinion polling. In quota sampling, quotas are pre-assigned to strata, but the final selection of cases is left to the discretion of the interviewer. However, if the benefits of random sampling are to be attained, subsamples from strata must be randomly chosen. The resort to convenience or judgment sampling is no more warranted within a given stratum than it is in the whole, unstratified population, and is almost certain to lead to biased results.

Since stratified sampling is more complex than simple random sampling, we may rightly ask what are its compensating advantages. Briefly put, it is a labor-saving device for securing equivalent accuracy with fewer cases than is likely under simple random sampling. It is essentially for this reason that national public opinion polls universally resort to some form of stratified sampling in order to keep the size of the sample down to manageable proportions. Thus, subsamples are usually selected from each of the broad geographical regions of the United States, it being plausibly assumed that public opinion on many issues varies from one sector to another: the Midwest, for example, is internationally more isolationist than the East. Similarly, subsamples are commonly drawn from various age and economic levels, as these factors are also known to exert an influence on the content and intensity of public opinion. Older persons are generally more conservative than younger; political opinions tend to parallel economic interests. On the other hand, national opinion polls would never stratify the population by hair color, since there is probably no genuine correlation between hair color and political opinion. In general, stratification serves no purpose when the stratifying factor is uncorrelated with the sampling trait being measured.

Comparison of Strata. While the principal justification of stratified sampling is equal accuracy with a smaller sample, the comparative data which are its natural by-product provide an additional inducement for using it. Thus, the decision to stratify by occupation in a survey of opinion on labor unions may be prompted as much by the wish to compare the characteristics of the various occupational classes as by the need to economize on sample size. The differences among occupational categories (Table 11.1) on the question "Are you in favor of labor unions?" may, in fact, be even more pertinent and revealing than the over-all weighted average of 67 per cent, which necessarily conceals such differences. Since strata are often treated individually, it has been suggested that the term "domains of study" be applied to strata when they are being analyzed in this segregated manner.

Table **11.1** *Attitude Toward Labor Unions by Occupation, Percentage Distribution, U.S.*

OCCUPATION	FAVORABLE	UNFAVORABLE	TOTAL
Farmers........	52%	48%	100%
Businessmen....	66	34	100
White Collar....	69	31	100
Professional.....	77	23	100
Skilled.........	75	25	100
Unskilled.......	71	29	100
TOTAL	67%	33%	100%

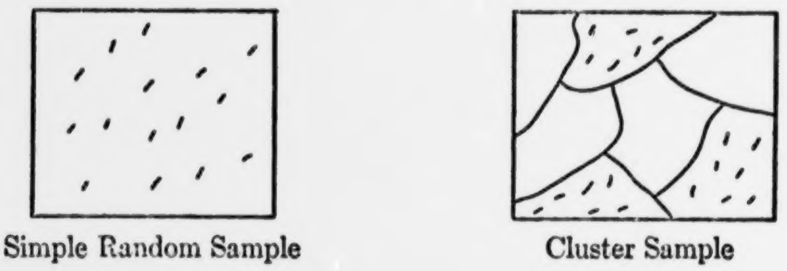
Source: American Institute of Public Opinion, May, 1942.

Cluster Sampling. In the case of the drawing for the door prize, it was obviously necessary to have a complete list of eligible names, from which the winning one was randomly drawn. For, "equal opportunity to be included" implies not only a mechanically reliable selection procedure, but also a complete list, so that every name is accessible to the draw. It would similarly be necessary to have a complete source list of sampling units if a simple random sample were drawn from any other statistical universe, such as the inhabitants of a community, an army, or a university student body. But such social universes differ from the more or less trivially small, easily manipulated guest list of a banquet. When the universe is vast, and extends over a wide area, the compilation of such an indispensable list is a laborious undertaking; and even when available (e.g., a city directory), it is usually not up-to-date on the date of issue because of the mutability of the population.

If, therefore, an adequate list of elementary sampling units is not available, we may turn instead to more or less permanent groupings into which the population is naturally divided and which can be conveniently listed. Human beings are, of course, usually found in prevailingly standard groups: they are clustered geographically by states, counties, municipalities, neighborhoods, blocks, precincts, and dwelling units; people work together in factories, offices, and stores; they are organized in innumerable clubs, lodges, schools, and miscellaneous associations. It is these groups, or *clusters*, which may be serially utilized as sampling units, through which we reach the ultimate elementary unit (e.g., the person or household) which is the objective of our survey. Such a procedure is therefore termed *cluster sampling.* We thus finally reach the elementary unit through a shorter or longer chain of samplings of the more easily listed clusters. Since sampling is carried out in successive stages before reaching our destination, this type of sampling is also called *multi-stage sampling.* When the clusters at any stage consist of territorial units, we may describe that stage as *area sampling.*

Let us suppose that we wish to survey the occupational ambitions of high school pupils, ages 15–17, in Chicago. It would be exceedingly laborious and financially prohibitive, as well as hazardous in accuracy, to compile a list of all specified pupils in that metropolis. However, a permanent list of high schools is easy to obtain. These schools could be randomly sampled, and a list of students in the desired age categories would then be obtained from the comparatively few high schools in the sample. From that list, the sample of students would be finally drawn. Similarly, if the households of the city were to be surveyed, a sample of city blocks, then of dwelling units, and finally of households would be drawn. Even though such a multi-stage procedure would still require a source list at each level, the lists would be smaller and more current. In such economy lies the first advantage of the cluster approach.

However, the execution of any social survey is not fulfilled by the mere drawing of the sampling units. We still must make the personal contact with the selected elementary units to obtain the information which was the objective of the study in the first place. This problem of personal contact develops into a major obstacle in simple random sampling, which may yield a sample whose elements are widely dispersed over a wide area, requiring a prohibitive expenditure of time and energy to reach. It is in such a plan for personal interviews that cluster sampling displays another technical advantage. Under this procedure, the elementary units are territorially concentrated, and are therefore more easily accessible, with less wastage in transportation. This is shown in Figure 11.1.


Simple Random Sample

Cluster Sample

FIGURE 11.1 *Schematic Diagram, Simple Random and Cluster Sampling*

But the physical and mechanical convenience of cluster sampling is purchased at a price in quality. It generally lacks, to a degree, the very characteristic which is the objective of good sampling: typicality and representativeness. Its reduced representativeness is due to the fact that the elementary sampling units within clusters, particularly human clusters, are likely to be closely similar in regard to their social characteristics; consequently, sampling within these clusters understates the dispersion and provides unnecessary duplication. For example, the residents of a given city block are likely to belong to the same socio-economic class, and therefore hold the same political and social opinions. By the same token, they are not likely to be representative of any diversity of public opinion in the large aggregate of unsampled neighborhoods; a cluster sample is thus *less* likely to represent the variety of opinions than a simple random sample of the same size scattered over the community. It is this lack of heterogeneity within clusters which reduces the comparative effectiveness of cluster samples.

This liability notwithstanding, cluster sampling is being increasingly employed in social research. For, after all, statistical work — like every other human endeavor — is a compromise between the ideal and the practical, and in any case can attain only an approximation of "truth."

Interval (Systematic) Sampling. Whenever the sampling units are arranged in some kind of natural sequence, like consecutive admissions to a hospital or library books in the card catalogue, it may be economical and even preferable to obtain a sample by taking cases at a fixed interval. Such a procedure is termed *interval sampling,* or more commonly but less aptly, *systematic sampling.* The selection of every tenth name from the telephone directory, after a random start among the first ten names, illustrates the process of interval sampling. Such a procedure has obvious utility for the social scientist who frequently has occasion to study a series of events such as a file of newspapers, the characteristics of dwelling units in a given ecological area, the cases on docket in a criminal court, or a card catalogue of welfare case records.

To establish the width of the sampling interval (k) in any given problem, we merely find the ratio of population size (N) to desired sample size (n):

$$k = \frac{N}{n}$$

Thus, if the sample is to contain 5 per cent of the universe, or 1 out of every 20 cases, the sampling interval obviously would be 20; and we would draw every 20th item, randomly starting with any number within the first interval of 20. Such calculation presupposes, of course, that sample size has been fixed in advance of the sampling. But if n has not been set, and an arbitrary interval is employed, it is still necessary to pass through the entire sequence, even though we may seem to have an ample number of items after we have proceeded only part of the way. If we discontinued the drawings before completing the entire circuit, we would deprive the units in the omitted segment of the opportunity of being chosen and thereby destroy the randomness of the operation. For example, by skipping names L to Z in an alphabetical listing, we would almost certainly produce a biased sample.

Since each interval contributes one and only one item to each sample, it follows that there can be no more different samples than there are items within the interval. Thus, if the sampling interval is equal to 10, there can be only 10 possible samples, regardless of the size of the universe, be it 1,000 or 1,000,000. But the number of different samples would be almost incalculable in simple unrestricted random sampling. This restriction in the number of possible samples serves to distinguish interval sampling from simple random sampling, since the latter furnishes an equal opportunity of selection to every distinct combinatorial sample of n items.

In spite of this severe limitation, interval sampling will often produce results that compare favorably in representativeness to those yielded by simple random sampling. In particular, whenever the values may be presumed to be randomly ordered, simple random and interval sampling will yield identically accurate results. Such a presumption is often reasonable when items have been alphabetically listed, since there is usually no correlation between the alphabetical order of names and the traits which the named objects possess. Thus, we have previously seen that the alphabetical listing of large American cities orders the respective suicide rates in a sequence which is seemingly purely random. Similarly, an alphabetical listing of students may be expected to result in a sequence of grade averages that is wholly random and, therefore, free of trends and cycles. In such cases, it makes no ultimate difference whether we select random digits or draw every *k*th unit, except that interval sampling is usually more simple to execute. The long-run results would be virtually identical for any given size sample.

There is, however, one notable circumstance that constitutes a special hazard for interval sampling and may easily lead to erroneous conclusions. When the universe values form a *cyclical progression*, the sampling interval may coincide with the phase of the cycle, causing interval sampling to yield an unrepresentative set of identical values. Let us consider a fictitious sequence whose phase is four: 1, 2, 3, 2; 1, 2, 3, 2; 1, 2, 3, 2. Now, if the sampling interval is set equal to 4, any sample will necessarily consist of a set of identical values. It will consist of all 1's, all 2's, or all 3's. Such samples will in no way do justice to the variation in the universe.

This type of pitfall is illustrated in a sampling study of June issues of the Sunday *New York Times*, 1932–1942, which disclosed that only Protestant marriages were featured on the society page of the sampled issues.* From this finding, the conclusion was drawn that the upper-upper social class of New York City was preponderantly Protestant in religious background. But this inference was immediately challenged on the ground that Jewish marriages happen for ceremonial reasons not to be performed in June, and therefore had no opportunity to appear in the issues of the *Times* which were sampled. A check sampling, undisturbed by daily and monthly cycles, revealed that Jewish marriages were in fact regularly featured by the society editors during the appropriate seasons. By that criterion, Jews were proportionately represented in the upper social strata. In this instance, the sample interval led to an overrepresentation of Protestants, an error compounded by the unfortunate judgmental selection of June as a point of origin.

But interval sampling also carries its intrinsic advantages. In fact, when the numerical values form an *arithmetic progression*, interval sampling will be even more effective than simple random sampling. For, in that event, the sample will necessarily distribute itself evenly over the entire range of values and thereby provide a reliable miniature of the population distribution. If, for example, we select every fifth boy from a lineup according to height, the resulting sample will necessarily be representative of the distribution of boys' heights. Analogously, if we sample every 25th dwelling unit along a metropolitan avenue after a random start, we would probably obtain an accurate cross-section of the various socioeconomic groupings on that street, since dwelling units are segregated and ordered according to social status. In this way interval sampling may supply its own stratification.

---

* David and Mary Hatch, "Criteria of Social Status as Derived from Marriage Announcements in the *New York Times*," *American Sociological Review*, XII, 1947, pp. 396–403; and W. J. Cahnman, "A Note on Marriage Announcements in the *New York Times*," *American Sociological Review*, XIII, 1948, pp. 96–97.

From this, it is evident that the principal advantage of interval sampling lies in the mechanical ease with which it can be applied to such natural sequences as rows of dwelling units, card files, city directories, and so on. Its special hazard is the cyclical sequence, and we must maneuver to circumvent it when that danger is thought to exist.

Interrelatedness of Random Sampling Procedures. Quite obviously, the foregoing random procedures are not mutually exclusive. They may be — and usually are — combined in a variety of ways. Cluster sampling may be used within broad strata, and interval sampling may be used within these clusters. No single *sample design* is best for any given purpose. In all sampling, an attempt is made to attain the desired degree of representativeness as economically as possible, which is the guiding criterion of modern sampling design. We should recognize that the designing of a sample is a form of statistical engineering and accounting, requiring appropriate skills and knowledge. In the foregoing statement, we have merely hinted at the technical aspects of sampling, which are of course fully developed in treatises on that subject.

However, effective sampling requires much more than mere technique. If the assets of a given sampling procedure are to be fully realized, it is essential that all necessary discretionary as well as mechanical steps be expertly performed. Thus, the anticipated benefits of stratified sampling will not be attained unless the strata into which the population was judgmentally divided before the sample was drawn actually differed among themselves on the sampled trait. In the aforementioned instance, stratification by race would be profitless if Negroes and whites shared the same opinions on school integration. If on the other hand, Negroes and whites differed widely, then, statistically speaking, there would have been considerable variation between strata, but relatively little within each stratum, thereby validating the original decision to stratify by race. It is more efficient in terms of sample size to sample from two homogeneous strata than from one very mixed, heterogeneous stratum. The knowledgeable worker must anticipate the validity of the stratifying criterion before the sampling begins.

Analogously, the effectiveness of cluster sampling will be enhanced by expertly composing and recomposing clusters in advance of the sampling so that each cluster is as representative of the entire population as possible. Insofar as that effort can achieve success, given the heterogeneity of the grand universe itself, the respective clusters will tend to resemble one another, while there will be considerable statistical variation within each cluster. Thus, in cluster sampling we invert the specifications of stratified sampling: instead of homogenizing strata we diversify the elements within clusters. We may, for example, combine precincts into larger geographic districts in order to increase the diversification within clusters and thereby raise their representativeness of the entire electorate and fulfill their function as samples.

In sum, no sampling technique is completely automatic; all involve subject-matter decisions. Hence, first-hand practical experience with the concrete subject matter contributes quite as much to fulfillment of a sampling project as dexterity in the mechanical routines of applied statistics.

QUESTIONS AND PROBLEMS

1. Define the following concepts:
    - Universe
    - Sample
    - Infinite Population
    - Finite Population
    - Sampled Population
    - Target Population
    - Availability Sample
    - Haphazard Sample
    - Judgment Sample
    - Random Sample
    - Simple Random Sample
    - Stratified Sample
    - Cluster Sample
    - Interval Sample
2. State difficulties that might be encountered in defining the populations:
    - Families
    - Dwelling Units
    - Farms
    - Households
    - City Blocks
    - Broken Homes
    - Overcrowded Homes
    - University Students
    - Gainfully Employed
    - Social Class

SURVEYS AND EVIDENCE

The following article illustrates the difficulties which lawyers interested in social science evidence will have to confront. Professor Zeisel tells us that there are two major problems related to using survey evidence in judicial proceedings. "The first problem raises the issue of hearsay. The second raises the question as to how interviewing a few hundred people . . . can produce correct conclusions concerning many hundreds of thousands of people." The second problem is essentially the same one addressed by Mueller and Schuessler in the preceding reading. As both they and Zeisel point out, the reliability of survey evidence is dependent upon how well the universe has been defined, the means by which the sample has been drawn, and the techniques used to carry out the actual interviews. These procedures are important simply because "courts generally have refused admittance to surveys in which they have discovered technical flaws."

But it is Professor Zeisel's discussion of the rule against hearsay evidence and how it relates to surveys that we feel is the most important part of the reading. The selection does not offer a guideline for practitioners interested in having surveys admitted into evidence, but it does provide us with a clear analysis of the legal profession's hesitancy in using techniques and procedures foreign to the traditional methods of resolving legal conflicts.

From Hans Zeisel, "The Uniqueness of Survey Research," Cornell Law Quarterly, 45 (1960) pp. 322-346.
c Copyright 1960 By Cornell University.
Reprinted by permission.

Hans Zeisel, "The Uniqueness of Survey Research,"
45 Cornell Law Quarterly, pp 322-346 (1960).

> *"Just when a scientific principle or discovery crosses the line between the experimental and demonstrable stages is difficult to define. Somewhere in this twilight zone the evidential force of the principle must be recognized . . . ."**

Barely a decade ago, in a trade-mark confusion case involving girdles for young ladies,[1] Judge Jerome Frank playfully supported his dissenting opinion by a survey which he himself called "not satisfactory." In fact, it was probably the worst survey ever made. He simply solicited his daughter's and her girl friends' opinions.

Nevertheless, Judge Frank had a point. Sadly remarking that "neither the trial judge nor any member of this court is (or resembles) a teen-age girl," he felt that where the opinions of teen-agers are at issue nothing but "information directly obtained" from them would suffice. Regretting that he had no "staff of investigators like those supplied to administrative agencies" he proceeded on the sound principle that any relevant information is better than none.

By now, the science of making surveys has come into its own, and with increasing frequency the courts are called upon to decide questions relating to the admissibility and use of survey evidence. Administrative agencies, not bound by the technical rules of evidence, will usually admit a survey and then evaluate its probative power.[2] In court trials, especially jury trials, the issue of a survey's worth will first arise when admissability is decided on. But whatever the formal occasion, the evidential value of surveys is at issue. This paper is an effort to contribute towards a better understanding of survey evidence, its value and its limitations, within the legal framework which the courts have begun to develop.[3]

---

† See Contributors' Section, Masthead, p. 347, for biographical data.

* Frye v. United States, 293 Fed. 1013, 1014 (D.C. Cir. 1923).

[1] Triangle Publications v. Rohrlich, 167 F.2d 959, 976 (2d Cir. 1948).

[2] . . . It has long been settled that the technical rules for the exclusion of evidence applicable in jury trials do not apply to proceedings before federal administrative agencies in the absence of a statutory requirement that such rules are to be observed. Opp Cotton Mills v. Administrator, 312 U.S. 126, 155 (1941).

[3] A review of the law and literature through 1953 will be found in Lester E. Waterbury, "Opinion Surveys in Civil Litigation," a paper delivered at the 1952 meeting of the American Association for Public Opinion Research, 17 Public Opinion Quarterly 71 (1953); a revised version in 44 Trade Mark Rep. 343 (1954); Note, 66 Harv. L. Rev. 498 (1953). A more recent and more detailed, but less analytical, inventory can be found in Barksdale, The Use of Survey Research Findings as Legal Evidence (1957). A review of the potentialities and shortcomings of public opinion research, important for both the lawyer and the survey practitioner, will be found in Blum and Kalven, "The Art of Opinion Research: A Lawyer's Appraisal of an Emerging Science," 24 U. Chi. L. Rev. 1 (1956). Appended (at 64-65)

The unique value of a survey consists in its being the only means of measuring the characteristics of a group where such a measure is at issue, *e.g.*, the proportion of people who confuse two trade-marks, the share of a market held by a competitor, the level of wages paid in an industry, the amount of timber burned in a forest fire, the proportion of items manufactured below specific standards.

The two survey aspects which raise major legal issues are: (1) whether the information is gathered through verbal statements made by the survey interviewees, or only through observation, counting or measuring by the field worker, and (2) whether the survey is based on a census enumeration involving *all* members of the group, or merely on a sample taken from that group.

The first problem raises the issue of hearsay. The second raises the question as to how interviewing a few hundred people, or examining a few specimens, can produce correct conclusions concerning many hundred thousands of people or a whole shipload of merchandise.

Accordingly, for our purpose, surveys may be classified into four groups: (1) *Census* surveys *not* involving *verbal statements, e.g.*, surveyors measuring the acreage of an area, and bookkeepers or accountants determining the amount of a designated type of expenditure; (2) *Census* surveys involving *verbal statements, e.g.*, the standard job of the U.S. Census; (3) *Sampling* surveys *not* involving *verbal statements, e.g.*, the job of the Food and Drug Administration in examining shipments of merchandise, a survey of car license plates of patrons to determine their geographic distribution,[4] and certain phases of accounting work;[5] and (4) *Sampling* surveys involving *verbal statements, e.g.*, public opinion polls and similar interviewing operations.

The major legal difficulties are compounded in surveys of type (4). These public opinion poll surveys, therefore, will form the center of this discussion.

The law has not yet developed a general rule for the treatment of all survey evidence, although it is recognized that only two types of surveys are clearly acceptable. Survey evidence produced by the U.S. Census, although it is hearsay and frequently based on sampling rather than "census" operations, is clearly admissible. The other type of survey is admissible, not because of its distinguished authorship, but because of the nature of the collected information. If the interview answers are not

---

are comments by two distinguished social scientists, Paul F. Lazarsfeld and Samuel A. Stouffer.

[4] This serves to establish the competitive position between retail outlets, or motion picture theatres. See note 73 infra.

[5] See infra p. 329.

gathered "for the truth of the matter asserted therein" the survey will be accommodated under the state-of-mind exception to the hearsay rule.[6]

Surveys which do not fall into either of these two categories can gain admittance only under the broader, less stringent approach (of which the state-of-mind exception is but a special case) which weighs the necessity of utilizing the evidence against the dangers of hearsay. Unless, therefore, a survey exception to the hearsay rule emerges at some future time, the admissibility of survey evidence, as a rule, will hinge upon the assessment of its worth.

In two recent cases, the sharp dividing line between state-of-mind surveys and other surveys was softened. In one, a state-of-mind survey was admitted because there seemed to be sufficient safeguards for admitting that particular hearsay evidence.[7] In the other, a survey was admitted "for the truth of the matter asserted" in spite of the fact that it *was* made, because the hearsay dangers were found to be negligible compared with the value of the promised information.[8]

The privileged status which the law accords to the operations of the United States Census provides an appropriate introduction for our discussion because here, in a curious way, the law has solved, expressly or by implication, all the major perplexities created by surveys and sampling operations. Thereafter, we will consider the problems of sampling and hearsay as they arise with respect to other surveys. Our discussion will suggest that the true difficulties with survey evidence do *not* arise from its being based on sampling operations or on hearsay. Rather, these factors will emerge as the legal justifications for avoiding difficulties that can probably be met more directly.

THE UNITED STATES CENSUS

Courts have not hesitated to take judicial notice of Census data, and some state statutes make such data competent "prima facie evidence" as a matter of law.[9] This is so although all Census data are technically hearsay, transferred as they are from the original respondent through a long chain of staff workers. Furthermore, much Census data are based on sample operations rather than a complete enumeration of all units.

One explanation for the fact that the sampling issue has never been raised in connection with Census data is, perhaps, that so few people know that many Census data are based only on samples. For instance, the Census income distribution for the United States population and its subgroups is based on the questioning of every fifth household head. Similarly, such household data as the "proportion of homes having mechanical refrigeration" are based on replies from inhabitants in every fifth dwelling unit. Yet the Census Bureau considers these findings not significantly less reliable than actual "census" data. It is an obvious waste of resources to ask a question of all respondents, if questioning every fifth will yield results of sufficient, if not greater, accuracy.[10]

---

[6] See note 37 infra.
[7] See note 39 infra.
[8] See note 40 infra.
[9] 5 Wigmore, Evidence § 1671 n. 11 (3rd ed. 1940).

Even if the sampling nature of some Census data were better known, it is improbable that it would be used to impeach their accuracy. This may be inferred from the emphatic court endorsement of the counting procedures of an official institution of much lesser standing. With respect to a sampling examination conducted by the Federal State Seed Laboratory in Alabama, a court said:

> Reports which are of a public nature and taken under competent authority to ascertain a matter of public interest are admissible in evidence against all the world.[11]

The survey operations of the Census have still another privilege vital to the success of any continuous interviewing operation. By statute, the information given to the Census enumerator is treated as privileged, which makes it impossible to verify the obtained information by calling interviewees as witnesses in court.

For survey operations not conducted by the U.S. Census or similar public bodies, all three objections—that they are based on samples, that they constitute hearsay, and that there is a reluctance to abandon the anonymity of their interviewees—provide serious obstacles for the law. One must ask, therefore, why a status given to the U.S. Census is being withheld from other survey operations? Whatever the legal formula, the true answer must lie in the confidence which the Census survey peculiarly commands. This confidence derives from the disinterested character of the Census operation and from the trust in its expertness.[12] It is sustained by the great simplicity and the factual character of most Census questions, which leave little room for error.[13]

---

[10] See note 25 infra.

[11] E. K. Hardison Seed Co. v. Jones, 149 F.2d 252, 257 (6th Cir. 1945).

[12] Compare, however, this passage in Judge Caffey's opinion in the Alcoa case:

> . . . I conceive of no reason for discriminating between private individuals and Government employees in formulating . . . what constitutes inadmissible hearsay.

United States v. Aluminum Co. of America, 35 F. Supp. 820 at 825 (S.D.N.Y. 1940).

[13] Still another sample survey which is admitted almost daily without much legal argument in our courts, and probably without awareness that it is a sample survey, is the mortality table, submitted as proof of the average life expectancy at any given age, subject to modification by special trials of the particular individual.

SAMPLING PRINCIPLES

Most surveys are based on samples because, practically speaking, there is no other way of conducting them. There are even a few situations where a complete census enumeration is not merely impracticable but impossible because the census would destroy the entire group to be measured. This is inevitable whenever the examination involves the destruction of the examined unit, as in most laboratory examinations of food and drugs.[14] Similarly, sampling naturally recommends itself whenever the universe (the technical term for the group from which a sample is taken) is either physically unavailable or remains undetermined as to its ultimate size. Estimating the value of burned timber is an example of the first;[15] estimating the potential loss of bus fare from a projected change-over to one-way traffic is illustrative of the second.[16]

The most frequent and, in the long run, most important advantage of samples is their overwhelming economy as compared to the often prohibitive costs of the census operation. While most universes would permit a census survey in theory, its costs, in terms of money and time, as a rule, are prohibitive. In an early attempt to approach something like a census survey in a Federal Trade Commission case, one of the commissioners voiced this eloquent protest:

> I want to register my protest at the way in which this case was conducted. About a thousand witnesses . . . were permitted to testify as to whether the use of the word 'Castile' when applied to a soap not made exclusively of olive oil, had the tendency to deceive the public . . . . [T]his piling up of cumulative evidence is an inexcusable outrage on the public . . . . The attorneys and the trial examiner traveled throughout the country for the purpose of taking the testimony of such witnesses. About 700 such were subpoenaed to testify at Spokane . . . . [T]estimony has caused the F.T.C. to waste hundreds of thousands of dollars.[17]

Judge Wyzanski, in the *Shoe Machinery* Case, notes that "If anti-trust trials are to be kept manageable, samples must be used."[18]

---

[14] . . . [a] test(s) . . . results in the article's being rendered useless . . . . [I]nspection and condemnation on the basis of samples tested is clearly contemplated by the Act [21 U.S.C.A. 334(c)].

United States v. 43 and one-half Gross Rubber Prophylactics, 65 F. Supp. 534, 536 (D. Minn. 1946).

[15] [The witness] took four sample plots in each 40 acre subdivision and estimated the destroyed merchantable timber . . . . We are not convinced that any improper elements entered into the estimates of the witness. . . .

Feather River Lumber Co. v. United States, 30 F.2d 642, 643 (9th Cir. 1929).

[16] The Court accepted a sample survey and the opinion evidence of an expert witness on the ground that it was "dictated by necessity, there being no other way to establish the probable loss of fares." Eighth Avenue Coach Corp. v. City of New York, 170 Misc. 243, 251, 10 N.Y.S.2d 170, 181 (Sup. Ct. N.Y. County 1939), aff'd, 286 N.Y. 84, 35 N.E.2d 907 (1941).

[17] From Commissioner Humphrey's dissent in Matter of James S. Kirk & Co., 12 F.T.C. 272, 289 (1928).

[18] United States v. United Shoe Machinery Corp., 110 F. Supp. 295, 305 (D. Mass.

Occasionally, a situation may arise where the universe to be examined is finite and small enough so that the law, by insisting upon complete enumeration, does not impose an impossible burden on the litigant. In a California case, involving suit for overpayment of a local sales tax, plaintiff offered an estimate of the overpayment based upon a sample, but the court insisted upon an introduction of the complete count.[19]

The courts generally are familiar with the fact that a sample must be representative of its universe. To state precisely what is meant by a representative sample, we might start by recalling a most famous instance of what was *not* a representative sample—the more than two-million ballots of the 1936 Literary Digest poll which forecast Landon's victory over Roosevelt. By relying on the voluntary response from people listed in telephone directories, owners of automobile licenses, and its own subscribers, the Literary Digest sample *under*-represented those social strata which were to sweep Roosevelt into his second term. A representative sample, then, is one in which all members of the group have an equal chance of being selected. Such a sample will represent all subgroups more or less proportionately.[20]

The only way to insure an adequate representation is by some form of lottery. A lottery, with *all* members of the universe susceptible of selection, is at the core of every sampling operation. The so-called systematic sample, whereby every *n*th item or name from a list is selected, is but a hidden form of lottery. It is implied that these *n*th units are in fact random selections.[21] The many variants and problems in sample design do not so much derive from the basic and easily solved problem of se-

---

1953), aff'd, 347 U.S. 521 (1954). Compare also the strong endorsement of survey evidence in "streamlining the Big Case" Report of Special Comm. of the Section of Anti-Trust Law, Am. Bar. Assoc., Sept. 15, 1958, p. 31.

[19] Sears Roebuck & Co. v. The City of Inglewood, reported in Sprowls, "The Admissibility of Sample Data into a Court of Law," 4 U.C.L.A.L. Rev. 222-32 (1957). See also note 34 infra.

[20] Sometimes, however, if one wants to measure characteristics of a relatively small subgroup it will be necessary to over-represent that sub-group in the sample (over-sampling) so as to have a sufficiently large cross-section of the sub-group. If one wanted to compare, for instance, households with and without automobiles, one might sample every non-owner but only every fourth owner, so as to have equally sized samples of both groups.

[21] Judge Wyzanski used a sample of similar design:

> At the Court's suggestion, the Government took and offered in addition to the OMR's [Outside Machine Reports] depositions [a sample] of 45 shoe manufacturers operating 55 factories. The Court arbitrarily selected from a standard directory of shoe manufacturers, the first 15 names that began with the first letter of the alphabet, the first 15 names that began with the eleventh letter of the alphabet, all 8 of the names that began with the twenty-first letter of the alphabet, and the first seven of the names that began with the twenty-second letter of the alphabet. This sample covers 3 per cent of the shoe manufacturers. The sample includes small and large factories, and concerns manufacturing shoes according to substantially the most popular shoe manufacturing processes.

United States v. United Shoe Machinery Corp., 110 F. Supp. 295, 305. (D. Mass. 1953).

Such clustering, however, is not without danger since it may coincide with natural clusters on the list, such as the names beginning with "Mc" or with "Rosen."

lecting a representative sample,[22] but from the secondary difficulty of selecting a sample with *maximum efficiency,* that is, with minimum costs.

LIMITATIONS UPON THE ACCURACY OF SAMPLE MEASUREMENTS

The price to be paid for accepting a sample instead of a census enumeration is a certain degree of uncertainty. This might seem a shocking admission in an argument for the acceptance of sampling evidence. But this shortcoming, as we shall see, is mitigated, first, by the fact that the degree of uncertainty can be reduced to any desired magnitude (though never to zero); second, because the law is rarely in a position to insist on certainty;[23] and third, and most important, by the fact that census operations, too, are open to error. Experts insist that a sample survey, in spite of its inherent uncertainty, can, at times, be more accurate than the corresponding census operation.[24]

Let us first clarify the nature of this uncertainty in sample measurements. Suppose we found that 68 per cent of the housewives in a sample had bought at the XX food store chain. What can we say about the corresponding percentage in the universe from which this sample was drawn? Ignoring the technical niceties, the relationship is expressed as follows: A range is set, extending in both directions around the sample measurement, *e.g.,* from 62 to 74 per cent, also written as 68 ± 6 per cent. The probability that the true (universe) measure, if determined by a census, will fall within this range is then determined. If in this particular instance the sample measurement of 68 per cent was derived from a sample of 5,400 housewives, the odds will be 99 to 1 that the statement is correct that the true universe percentage will fall within the indicated limit. Given a certain sample, the odds will be lower where the narrower limits of the range are posted. In our example, the odds will be about 19 to 1 that one is correct in asserting that the true value will fall within 68 ± 4 per cent, and only 2 to 1 that they will fall within 68 ± 2 per cent, *i.e.*, between 66 and 70 per cent.

---

[22] The Federal Seed Act Regulations, 7 C.F.R. §§ 201.208-19, and especially §§ 201.210-13 (1959) contain explicit sampling instructions and thereby acknowledge this underlying principle.

[23]
> . . . [A]lthough, admittedly such summaries are not, and could not be, absolutely complete, they are the most accurate and most dependable data in existence to establish this fact.

Hughes Tool Co. v. Ford, 114 F. Supp. 525, 542 (E.D. Okla. 1953).

[24] The paradox is easily explained. Census operations are, by definition, gigantic tasks, necessitating the hiring of a great staff, the quality of which eludes control. It is suggested that population sampling surveys, in certain respects, would produce more accurate results than the population 'census.' Some of the Census inaccuracies are well known: Reported age distributions have a slight tendency to bulge at the round (10 year) intervals; transient (hotel, etc.) populations are underrated; babies under 1 year similarly are underrated—at the expense of babies who are reported as having reached their first year.

For the array of the survey problems the Bureau of the Census is confronted with, compare Hauser, "Labor Force and Gainful Workers—Concept, Measurement, and Comparability," 54 J. Sociology 338 (1949).

Compare, also, the following reference from a case before the Illinois Commerce Commission:

> The Company's evidence with respect to depreciation . . . was based primarily upon a sample of the Company's property . . . . It was pointed out that the use of a sample tends toward greater accuracy in the final conclusions, inasmuch as a reasonably intensive inspection of the entire plant is virtually impossible considering the limitation of time and money that should reasonably be imposed.

Illinois Bell Telephone Co., 7 P.U.R.3d 41, 606 Ill. Commerce Comm'n 493, 506 (1955).

A special case of this estimating procedure is acceptance sampling, a customary practice in quality control of production and accounting. The task is to determine, by means of a sample and in terms of agreed-upon odds, whether the proportion of faulty items in the universe does or does not exceed a specified tolerable minimum. If no faulty item is tolerable, sampling must, of course, never be resorted to, as that item might be hidden in the non-sampled part of the universe.[25] But if some small proportion of faulty items is tolerable, however small, a sample can be designed to assure confidence odds for such a margin.[26]

This situation requires the survey planner to determine the necessary accuracy of the sample measurement and to design his sample accordingly. The accuracy with which a sample result can be projected depends, as a rule, on two factors: the *absolute size* of the sample and the *variance* of the property to be measured. The widely held belief that the accuracy of a sample is connected with its relative size to the universe is mistaken.[27] A sample smaller than 1 per cent, taken from one universe, can be much more reliable than one comprising 10 per cent of another. To determine with equal accuracy the average age of the population of New York City and of Peoria, Illinois will require samples of equal size.

But if the *size* of the universe is immaterial, its *variance* is of great importance. Assume we wanted to determine the mean age of two groups of people and wanted to be sure of certain minimum odds that the sample result be not more than 3 years off the true mean age of the sampled group. Suppose, further, that one group to be sampled is the *general* population of a city and the second is the residents in all the *old age homes* of that city. Since age in the general population varies from zero to approximately one-hundred, the general population will, for the same degree of accuracy, require a larger sample than the old age home population, whose age varies only between seventy and one-hundred.

---

[25] To think otherwise would be like the person who, in counting a bundle said to contain 1000 one dollar notes, stops satisfied at 996 and exclaims: "If it was all right up to now, the rest will be all right too."

[26] Sampling, as a legitimate part of the accountant's practice, is acknowledged in Ultramares Corp. v. Touche, 255 N.Y. 170, 191-92, 174 N.E. 441, 449 (1931).

> The defendants [charged with having falsely certified a balance sheet] attempt to excuse the omission of an inspection of the invoices proved to be fictitious by invoking a practice known as that of testing and sampling. A random choice of accounts is made from the total number on the books, and these, if found to be regular . . . are taken as a fair indication of the quality of the mass. . . . Verification by . . . sample was very likely a sufficient audit as to accounts regularly entered upon the books in the usual course of business.

The court goes on to say that the sample in the present case did not excuse the omission, not because sampling in itself is improper, but only because the crucial entry in the ledger of assets, plainly interpolated and suspicious on its face, was excluded from the sample. For technical details on sampling in accounting work, see Vance and Netter, Statistical Sampling for Auditors and Accountants (1956).

[27] Only in those rare cases where the sample becomes a sizable proportion of the universe, say 20 per cent or more, do adjustments for approaching the size of the universe need to be made in the formula.

The inverse relationship between variance and sample size is most visible in the sampling of well mixed liquids; a drop of blood will provide an accurate count of the blood cells per volume, because this percentage varies little from drop to drop. There is, indeed, no more homogenous universe than a well mixed liquid. Hence, even a very small sample might yield an accurate measurement. Yet the following generalization is not warranted:

> [One of the] prerequisites necessary to the admission in evidence of samples [is] that the mass should be substantially uniform with reference to the quality in question. . . .[28]

The court errs here, for also a universe with low homogeneity permits sample measurements of great accuracy; it merely requires a correspondingly larger sample. Uniformity or homogeneity or variance—the three terms are used here interchangeably—is a matter of degree and can be compensated by appropriate sample design and size.

Two points emerge; first, that the degree of accuracy of any sample can be increased to the desired limit; and second, that the degree of accuracy with which the sample is related to its corresponding value in the universe can be stated with precision. Expert knowledge of the sampling error is as unambiguous as the knowledge of the relationship between a circle's radius and its circumference.

The Sampling Error and the Law

Sample measurements raise interesting legal issues. It might seem disconcerting that no sample measurement can be stated with complete confidence in its accuracy. But, as we pointed out, the law is accustomed to dealing with less than perfect evidence. Both standards of proof, "reasonable doubt" and "preponderance of evidence," allow for imperfections. Rather, it is the positive aspect of sample measurements that raises new issues, *i.e.*, the possibility of actually measuring the degree of uncertainty or certainty through tolerance limits for the measurement at issue.

One way in which the law might deal with the problem is to accept the range of the sampling error because *any* value within its range would fulfill the immediate legal requirements:

---

[28] E. K. Hardison Seed Co. v. Jones, supra note 11 at 256.

> For the evidence is all in one direction . . . . Different methods of weighting the various factors would produce different results; but no reasonable, qualified person would, by any rational process, reach a figure outside that range; and probably most methods would reach a figure close to the middle of it. Moreover, even though this high figure is not attained in every part of the market, nonetheless the figure may be fairly used since United supplies in every significant generic class of shoe machinery, except machinery used in the cement process; and in rubber shoe manufacturing, and of course, excepting dry thread sewing machinery, far more than 50% of the demand. In short, it is not inaccurate in this market to say United has a 75-95% share; and it probably would be accurate to say an approximately 85% share.[29]

Or the result of the measurement, however inaccurate, may clearly fall short of the legal requirement:

> . . . [P]laintiff did produce witnesses who testified as to confusion . . . . However, considering that annually more than 70,000,000, rolls of plaintiff's . . . mints are sold, to say nothing of many millions of packages of other flavors . . . it would be extraordinary if some confusion could not be found irrespective of the details of the dress of the package . . . . A new competitor is not held to the obligations of an insurer against all possible confusion.[30]

In these cases the court found the magnitude of the sample measurement so clearly above or below the relevant legal limits that it deemed it unnecessary to put numerical values on these limits. In the following opinion, however, full use was made of the possibility of measuring the accuracy of a sample:

> The Commission retained [an expert] . . . to analyze the Company's sample of its plant to determine . . . whether it was fairly representative of the plant as a whole. He testified that in his opinion the sample was fairly drawn and that the chance of it varying as much as 1% from the plant as a whole was negligible. We conclude that the Company's sample should be accepted as a fair cross-section of the plant as a whole.[31]

The courts could go further and specify standards for sample measurements since the sampling error can be reduced at will by enlarging the sample. To be sure, there is a serious cost element to be noted. Since the size of the sample error is inversely related to the *square* of the sample size, the size of the sample must be quadrupled to reduce a sampling error to one-half of its original size.

An interesting solution was proposed in the California case in which suit was instituted for overpayment of a local sales tax. Plaintiff offered an estimate of the overpayment based on a sample survey as follows: $28,250 with 2 to 1 odds that the true (census) value would fall between $23,150 and $30,350, and 19 to 1 odds that it would fall between $23,950 and $32,450. Plaintiff voluntarily reduced his claim to $27,000, thereby suggesting his willingness to absorb the greater part of the sampling error. The court insisted on a complete census count, however, only to discover that it yielded $26,750, a sum within 1 per cent of the claimed amount.[32]

---

[29] Judge Wyzanski in United States v. United Shoe Machinery Corp., supra note 21 at 307.

[30] Life Savers Corp. v. Curtiss Candy Co., 182 F.2d 4, 8 (7th Cir. 1950).

[31] Illinois Bell Telephone Co., supra note 24.

In most situations, however, the universe to be measured will be so large that the court will not have the choice between a sample measurement and a census, but rather, would either have to accept a sample measurement or have no measurement at all. In the latter situation the relatively inaccurate sample measurement may provide better evidence than the law now possesses. That a sample value may be inaccurate by some small margin of error could be of less importance than the fact that it *can* provide a measurement as accurate as it does. This possibility becomes particularly important when the measurement refers to the core of the litigated issue, and hence becomes a measure of the soundness of the court's judgment or verdict. Two examples may be cited.

The owner of a gambling establishment, at his trial[33] for income tax evasion, contended that his lottery wheel had retained, over the years in question, some 11 per cent of the placed bets. The prosecution introduced expert evidence to the effect that the laws of probability suggest that the wheel (with its specific game rules) must earn, in the long run, some 22 per cent of the gross intake, that is, about twice the amount for which the defendant argued. The prosecution might have gone one interesting step further and offered evidence in the following form: "The odds that the defendant's statement is true—that the wheel earned not more than 11 per cent—are one in a thousand" (or whatever the odds were, computed from the rules of the game and the actual number and size of the placed bets).

The use of blood tests in paternity suits provides another example. It appears that, because of the possibility of accidental mutations, their evidentiary value is not completely foolproof. On the average, it is estimated that the blood test may err, by excluding the actual father from parenthood, in one birth in 10,000. Because the paternity cases which reach the courts are more likely to be among the few where the test fails, it is estimated that the test would lead the court into error about 1 time in a hundred. The question arises, therefore, whether, in view of this fallibility, the blood test can still be accepted as incontrovertible proof. The Danish jurist who drew attention to this particular aspect of the problem suggests that blood tests be accepted as such proof. To permit their impeachment by witness testimony could only increase a ratio of error that is otherwise kept to 1 in 100.[34]

---

[32] Sears Roebuck & Co. v. The City of Inglewood, supra note 19. The issue posed in that case could provide a more general rationale for deciding the degree of accuracy that should be required of a sample. The answer could come from asking these two questions: (1) How much would it cost to increase the accuracy of the sample to a specified limit? (2) What would be the dollar equivalent of this gain in accuracy for the issue at hand?

[33] United States v. Sanders Scott, 55, 113 (7th Cir.) (unpublished).

THE HEARSAY BARRIER

Since opinion surveys reflect statements made by third persons to an interviewer who, in turn, has related them to the analyst who may appear as an expert witness in court, such evidence is clearly hearsay. But the courts have developed a line of authority which exempts surveys from the hearsay rule if the declarations are reported not "for the truth of the matter asserted therein," but as expressions of the interviewee's state-of-mind. Confusion of names and trade-marks provide the leading examples:

> The hearsay objection is unfounded . . . . [T]he statements of the persons interviewed were not offered for the truthfulness of their assertions . . . [T]hey were offered solely to show as a fact the reaction of . . . the public. . . . Only the credibility of those who took the statements was involved, and they were before the court.[35]

It is true that some of the hearsay dangers are reduced by an approach of this kind; but they are by no means absent. Thus, while one may be inclined to welcome any opening for the acceptance of survey evidence, one cannot wholeheartedly endorse this particular one.

The state-of-mind doctrine has two disadvantages. It may make the courts overlook technical pitfalls which are more likely to occur in state-of-mind surveys than in other types of surveys. For example, the degree of confusion of two trade-marks is, as a rule, the sum of two forms of confusion: the specific confusion of the two trade-marks in question plus the general confusion that will obtain for *any* comparison in an average group of people under most test conditions. Clearly, unless the degree of specific confusion is great, allowance must be made for the general.

---

[34] Ross, "The Value of Blood Tests as Evidence in Paternity Cases," 71 Harv. L. Rev. 466, 483 (1958). He adds, significantly, that the courts in Denmark are inclined to accept the blood-test exclusion as absolute proof with respect to children born out of wedlock, but show greater reluctance in cases where the paternity of legitimate children is at issue.

[35] United States v. 88 Cases, 187 F.2d 967, 974 (3rd Cir. 1951), cert. denied, 342 U.S. 861 (1951). See also People v. Franklin Nat'l Bank, 200 Misc. 557, 105 N.Y.S.2d 81 (Sup. Ct. Nassau County 1951), rev'd, 305 N.Y. 453, 113 N.E.2d 796 (1953), rev'd 347 U.S. 373 (1954); Household Finance Corp. v. Federal Finance Corp., 105 F. Supp. 164 (D. Ariz. 1952). While public opinion polls become almost standard procedure in trade-mark confusion cases, curiously enough, they have never been used to prove (or disprove) the assertion that a trade name had become a generic term.

The more serious difficulty with the state-of-mind rule, however, is that it provides no basis for admitting surveys of *facts*. In the latter type survey, the hearsay danger is less than in surveys of states-of-mind. Surveys which ask the respondents whether they own a gas stove, or carry a certain brand of merchandise in their store, or have reached a certain age, are not admissible under this hearsay exception. Surveys asking for complicated psychological reactions *are* covered. Judge Wyzanski has noted this problem and has decided against the mechanical application of the state-of-mind rule. After reviewing the positions and citing authorities for each,[36] he concludes:

> . . . So long as the interviewees are not cross-examined, there is no testing of their sincerity, narrative ability, perception, and memory. There is no showing whether they were influenced by leading questions, the environment in which questions were asked, or the personality of the investigator. But where a court is persuaded that in a particular case all these risks have been minimized, that the answers given by the interviewees are, on the whole, likely to be reliable indicia of their states of mind, that the absence of cross-examination is not prejudicial, and that other ways of getting evidence on the same point are either impractical or burdensome, the testimony should be admitted. . . . In this case these conditions have been met. Accordingly, the hearsay objection is overruled and the testimony of the results of the poll is admitted. . . .[37]

While this position considerably narrows the state-of-mind exception applicable to survey evidence, it must, in the long run, prove sound.

Another case which provides an opening wedge for surveys which cannot be accommodated under the state-of-mind exception is *United States v. Aluminum Company of America*.[38] There, an expert testified to an analysis of the records of 605 test holes drilled to determine the presence and quality of bauxite deposits, although the persons who did the drilling and prepared the reports were not present as witnesses:

> Opinion testimony by an acceptable expert resting wholly or partly on information, oral or documentary, recited by him as gathered from others, which is trustworthy and which is practically unobtainable by other means, is competent even though the first hand sources from which the information came be not produced in court.[39]

---

[36] Some authorities have, therefore, concluded that the testimony is not hearsay. United States v. 88 Cases, 187 F.2d 967 (3d Cir. 1951); 6 Wigmore, Evidence (3d ed.) § 1776; Note 66 Harv. L. Rev. 498, 501, 503, note 34. Others, however, have noted that the proffered evidence has some of the dangers of hearsay. See Note 66 Harv. L. Rev. 498, 501-502; Morgan, Hearsay Dangers and the Application of the Hearsay Concept, 62 Harv. L. Rev. 177, 185, 202-203, 206; McCormick, The Borderland of Hearsay, 39 Yale L.J. 489, 491.

American Luggage Works v. United States Trunk Co., 158 F. Supp. 50, 53 (D. Mass. 1957).

[37] 158 F. Supp. at 53.

[38] 35 F. Supp. 820 (S.D.N.Y. 1940).

[39] Id. at 823.

The court cites as precedent the opinion of Judge Learned Hand in *G. & C. Merriam Co. v. Syndicate Publishing Company*:[40]

> With respect to the matter, in what impresses me as unambiguous authoritative judicial language, it has been said that 'the requisites of an exception of the hearsay rule' are 'necessity and circumstantial guaranty of trustworthiness.'

The court goes on to clarify the term necessity:

> In effect, . . . necessity . . . is not to be interpreted as uniformly demanding a showing of total inaccessibility of firsthand evidence . . . but that necessity exists where otherwise great practical inconvenience would be experienced in making the desired proof. . . .[41]

The opinion, standing alone, is limited. The information in question was not gathered through interviews; hence it involves only one level of hearsay; moreover, it had been gathered in the "ordinary course of business" and only later used for purposes of litigation. Nor does the decision refer directly to surveys but to expert testimony informed *by* a survey. It could, however, be extended to the survey itself.[42]

Let us now consider more closely the dangers of insincerity, faulty narration, perception and memory as they pertain to survey evidence. But let us be sure to see the problem in its precise form: the issue is not whether the reliability of interview response would increase if all interviewees could be examined as witnesses in court, since this is not an available alternative. As a rule, it is not possible to bring the universe or its truly representative sample into court. The customary procedure is to call a number of public witnesses who allegedly are representative of the universe. But a distinguished lawyer with broad experience in this field had this to say about such a procedure: "The poisonous feature of the public witness matter is . . . that all too frequently they are selected not impartially but because they will testify the way the party selecting them wants them to testify."[43] The very fact that such witnesses are arbitrarily selected should render their testimony less credible. Even if these public witnesses would, in fact, give a more reliable response in court than to a survey interviewer, their evidence should be rejected on the ground that these individuals do not adequately represent the universe,[44] no matter how many of them are called.[45]

---

[40] 207 Fed. 515, 518 (2d Cir. 1913).

[41] Supra note 38 at 823-24.

[42] . . . [I]t may very well turn out to be that when the admissibility of survey evidence is finally established beyond doubt, the rationale will be taken from the field of expert testimony.

Waterbury, "Opinion Surveys in Civil Litigation," 44 Trademark Rep. 343, 362 (1954).

[43] Id. at 347 Waterbury discusses this practice in detail. He notes its appearance in Stanley Laboratories, Inc. v. FTC, 138 F.2d 388 (9th Cir. 1943), and Book-of-the-Month Club, Inc. v. FTC, 202 F.2d 486 (2d Cir. 1953).

Assuming it were possible to bring a truly representative sample of public witnesses into court, it is doubtful that their answers would always be more reliable than those given to a survey interviewer. The interviewee is in no way connected with the litigants, not even through the tenuous bonds created by being a witness for one side. Moreover, the interviewee will, as a rule, not learn the purpose for which his response is used. In a proper survey routine, to prevent inadvertent disclosure, not even the interviewer is told of the survey's purpose.[46] In addition, since the survey necessarily precedes the trial, less time will have elapsed between the response and the event to be recalled, than between the event and its deposition at the trial. Finally, the court has before it the complete and uniform question schedule in response to which the survey results were obtained. Court witnesses, on the other hand, at times undergo careful individual preparation prior to trial, the form of which does not necessarily come to the court's knowledge.

To be sure, court witnesses may have a heightened awareness of what is at issue and may be more careful and perhaps more perceptive than survey respondents who are completely unaware of the ultimate issues.[47] Cross-examination, too, may prove its value at any time that recollection or narration proves faulty. But even if some of the *individual* survey responses are not, in fact, as totally accurate,[48] the group measurement may still be sufficiently accurate within set tolerance limits.

In summary, therefore, the advantages offered by survey responses should at least suffice to protect such evidence from outright disqualification as hearsay. Moreover, the questions propounded in many cases will be so simple, straight-forward and unambiguous that the hearsay dangers must, in fact, be negligible. Hence, survey evidence, if properly procured, is well covered by Wigmore's formulation of the rationale that underlies all exceptions to the hearsay rule: "Where circumstances are such that a sincere and accurate statement would naturally be uttered, and no plan of falsification be formed. . . ."[49]

---

[44] The universe, of course, might be so small and accessible as to permit examination in court of every one of its members.

[45] Hence, it is wrong to believe that the evil of the procedure could be cured by calling a sufficient number of witnesses ". . . except by calling as witnesses so many of the public as to render the task impracticable." People v. Franklin Nat'l Bank, supra note 35 at 566. A badly selected sample only becomes worse as it becomes larger. See note 64 infra.

[46] "[The] witnesses were not informed of the purpose of this employment [a shopping survey]. . . . Always two went together so that there were two witnesses to each sale." Oneida, Ltd. v. National Silver Co., 25 N.Y.S.2d 271, 286 (Sup. Ct. Madison County 1940).

[47] "Purchases of merchandise are not made in a vacuum with Professor Quiz in charge." Quaker Oats Co. v. General Mills Inc., 134 F.2d 429, 433 (7th Cir. 1943).

[48] Hence, Judge Arnold's concern in New York Life Ins. Co. v. Taylor, 147 F.2d 297, 304 (D.C. Cir. 1944) is not quite to the point:

> A corporation is engaged in taking a nationwide poll as to the number of members of the Communist Party. In the regular course of that business . . . the interviewer reports that X, Y and Z are Communists, giving excerpts from the conversations to support this opinion. The report would [not] be admissible . . . to make a prima facie case that X, Y and Z are Communists . . . Such evidence might be used in a proceeding for the cancellation of a naturalization certificate.

To be sure, if the unreliability of the individual response passes a certain point, it may invalidate a survey.

SURVEY INTERVIEWEES AS WITNESSES

Some court decisions suggest that a right of the adversary to call survey interviewees as witnesses might help to overcome the hearsay obstacle by fortifying the reliability of the survey evidence.[50] Although in some cases this procedure has proved feasible, and might seem advisable in others, there are strong reasons against making it a general requirement for the acceptance of surveys. Cross-examination of selected survey interviewees is likely to be misleading. It may lend the aura of reliability to an incompetent survey, or it may destroy confidence in a survey which deserves better. Careful studies have shown that, on reinterviewing, one always finds respondents who change their original response.[51] These studies have also shown, however, that such changes, as a rule, do not affect the reliability of the survey. This paradox requires explaining although such explanation might, in fact, encourage the very practice against which it is aimed.

A change of response may be due to any number of causes, *e.g.*, simple response error during either the first or second interview. Even on such factual items as age, a small but definite group of respondents, on reinterviewing, will give a different response. Moreover, changes may occur which reflect opinions and attitudes which, in fact, may have been altered between the two interviews. Finally, the very fact of having been previously interviewed, may precipitate the change. The artificial stimulus of the interview may generate subsequent conversations or inquiries on the part of the respondent which, in turn, will affect the original uncontaminated attitude. A subpoena and what follows may prove even more disturbing; before the cross-examination is finished, so many influences will have become operative since the first interview, that a discovered discrepancy would prove little or nothing. In such a case all the weight would have to be accorded to the first response.

---

[49] 5 Wigmore, Evidence § 1422 (3d ed. 1940).

[50] ". . . nor any of the persons allegedly interviewed were called as witnesses. . . ." General Dry Batteries, Inc. v. Ray-O-Vac Co., 45 Trademark Rep. 588, 594 (1955)—"Any information he [the field representative] could give on the witness stand, would . . . have amounted to hearsay based upon hearsay." Irvin v. State, 66 So. 2d 288, 291 (Fla. 1953). The assurance that the interviewees could be called as witnesses was made a condition of the admission of the survey in Everitt Hat Company v. Solcum Hat Company, Milwaukee County, Circuit Court, July 11, 1955 (Wisc.), unreported opinion, quoted in Barksdale, supra note 3. "The defendant did not attempt to fortify its survey through any such witnesses." Oneida v. National Silver Co., supra note 46 at 287. Defendant subpoenaed 17 witnesses who had signed the survey. Quaker Oats Co. v. General Mills, Inc., supra note 47 at 431.

[51] Kendall, Conflict and Mood; Factors Affecting Stability of Response (1954).

One might argue that there is some value in cross-examining a respondent as to the circumstances of his being questioned and the mode of recording his answers. But these same circumstances can also be elicited from a cross-examination of the interviewers.

The discovery of discrepancies on the individual level need not invalidate the survey results, for while some interviewees may switch from position *a* to *b*, an approximately equal number will switch from position *b* to *a*. The difficulty lies in the fact that if such occasional individual changes are revealed in court through cross-examination, it may tend to impeach the survey because the over-all compensating effect cannot be shown.[52]

There is still another reason against encouraging this practice. Unless the courts protect the survey interviewees from subsequently being called as trial witnesses, bona fide surveys for purposes of settling legal issues will become more and more difficult to conduct. The problem is engendered not so much by the potentially large number of such witnesses, but primarily by a peculiar condition under which surveys operate. Interviewers of private survey organizations are finding open doors and willing respondents because they are scrupulously observing a canon which they share with the Census interviewer. They will not voluntarily identify individual answers by exposing their respondent. This assurance is given either explicitly or implicitly by all reputable survey organizations. Their operations would come to a halt if it were known that an interviewee might have to pay for his cooperation by being called into court and there exposed to the doubtful pleasures of cross-examination. The law specifically prohibits any disclosure of the answers of the respondent in Census surveys in order to insure truthful response and to avoid embarrassment. No other survey response enjoys this protection. A private survey organization cannot assure its interviewees of confidential treatment if the court orders the production of its survey questionnaires.[53] The knowledge of this threat keeps many a survey organization from accepting such legal work.

---

[52] In cases of major importance, the survey organization may attempt to document this by re-interviewing part of the respondents and comparing the paired results.

[53] The issue arose before an FCC examiner concerning the sale of station WGIS. It also arose during the hearings before an examiner for the U.S. Department of Labor concerning a survey of wages paid in the electric lamp industry.

There is a method of avoiding this dilemma.[54] The identification of the survey respondent can be written on a perforated section of the questionnaire and separated after the supervisor has passed it as satisfactory. The separated names will form a record of all respondents, but since none can be identified with any specific response, they can be cross-examined only as to whether they had been interviewed at all. The mutilation of questionnaires, of course, might be misinterpreted, but a proper explanation should eliminate this danger.[55]

For these reasons, then, the cross-examination of the survey respondents should *not* be required as a rule. In the vast majority of cases it will not aid in the evaluation of the survey but might, on the contrary, only confuse the issue. Whether the law, so hesitant to extend the area of privileged communications, will soon respond to these needs is doubtful.

THE IMPEACHMENT OF SURVEYS

The discussion thus far has tended to establish that neither the fact that a survey is based on a sample, nor that it relates hearsay evidence should, in itself, bar its admission into evidence. To be sure, there is the danger that the weakness of survey evidence may be hidden under a pretending surface. The following discussion will suggest that under the guidance of expert witnesses any such defects can be satisfactorily exposed and, hence, most danger avoided. The courts generally have refused admittance to surveys in which they have discovered technical flaws, rather than admitting them and permitting their impeachment. Administrative agencies, on the other hand, are more liberal in their practice, partly because they are not bound by the ordinary rules of evidence, and partly as a result of their familiarity with the specific technical problems before them. Moreover, many cases before these agencies are of the sort that can hardly be decided without the assistance of survey evidence.

A court sitting without a jury will seldom hesitate to admit a survey in evidence. The Supreme Court has never either reversed or criticized a trial court for admitting survey evidence in a civil case tried without a jury.[56] The court may indeed admit a survey even if it has no confidence in its probative value:

---

[54] It was employed in a survey by the Federal Food and Drug Administration in United States v. 353 cases *** Mountain Mineral Valley Water, Civil No. 565, E.D. Ark., May 1956.

[55] In a survey designed for an FCC hearing, the interviews were concluded with the following statement:

> As you may know, reputable survey agencies never make known any individual's opinion without his permission. The sponsor of this survey is seeking an application to operate the service we've been discussing with you. Would you be willing to have your name referred to him in connection with the opinions you have just expressed?

This approach is straightforward, but must lead to self-selection of those respondents who are more likely to stand by their outspoken opinions. The result must be a biased group of witnesses of no value to the court.

> It is doubtful that such an exhibit [a survey of farm machinery outlets in Iowa, conducted by the Statistical Laboratory of Iowa State College] has any relevancy . . . . [B]ut the Court concludes that there is no need to strike the exhibit. It may remain in evidence for what it may be worth. However, it should be stated that the conclusion hereinafter indicated . . . would be the same in absence of this documentary evidence.[57]

When sitting with juries, however, the courts prefer to exclude surveys if, on preliminary examination, they find flaws in them, rather than admit them for whatever they may be worth. Where the line between exclusion and admittance ought to be drawn should depend upon how difficult it is effectively to impeach a bad survey.

Improper Universe

There are three critical points at which a survey operation can fail and provide ground for its impeachment. First, the survey may have been directed at a universe which is irrelevant to the litigated issue. In such cases of obvious error, the court will not need expert advice.

> . . . [I]nterviewers stopped [the respondents] in front of one of the appellant's stores in San Francisco and asked them in what manner they spoke of 'Lerner Shops.' Obviously the results of such a survey are of little value in determining what knowledge residents of San Jose had of 'Lerner Shops' . . . .[58]

Or,

> . . . [T]he survey, having been limited to retailers, is inadmissible to show that in the market of ultimate consumers the plaintiff's design had acquired a secondary meaning.[59]

Inadequate Sample

It is not always obvious that a survey reflects an improper universe. Sometimes it purports to represent the correct universe but, in fact, does not do so. This is the second point at which a survey may prove inadequate. The universe may be properly selected, but the sample designed to represent it may be faulty. At this point the expert's help, as a rule, will be needed to explain the magnitude of unavoidable flaws or, as the case may be, of any errors in sampling.

The quality of any sampling procedure depends both on its basic design and its execution. The expert will readily discover its deficiencies from the report itself, supplemented by internal documents and such testimony from the survey staff as may be necessary. Questions directed at discovering the manner in which a respondent was selected for interviewing should provide all the information an expert will need. The survey staff, from the director down to the field interviewers, must be available for cross-examination. While cross-examination of *all* interviewers should be avoided, the court should not refuse to hear as many as are needed to clarify the exact modalities under which the survey was conducted.

---

[56] United States v. United Shoe Machinery Corp., supra note 21.

[57] United States v. J. I. Case Co., 101 F. Supp. 856, 858 (D. Minn. 1951).

[58] Lerner Stores v. Lerner, 162 F.2d 160, 162 (9th Cir. 1947).

[59] American Luggage Works v. United States Trunk Co., 158 F. Supp. 50, 52 (D. Mass. 1957).

To detect deviations from instruction will require a more detailed probing, primarily by questioning the supervisory staff and randomly selected interviewers.[60] Questions should be directed at the institutional safeguards against error (substituting, without permission, respondent *B* for respondent *A*) and against the admittedly rare occurrence of faking parts or the whole of an interview. These safeguards may include proper recruitment, training, and supervision of the field staff, as well as spot controls and double checks of the particular survey sample.

One of the more easily overlooked sampling traps may arise from what is technically called non-response. There are always some individuals in any sample from whom it is impossible to obtain the desired information, either because they could not be located (*e.g.*, were not at home) or because they refused to answer the questions asked of them. An effort to measure the size of broadcast audiences, for example, must go far astray if it bases its findings only upon the people found at home. It will exaggerate the true audience because such people are more likely to listen to broadcasts than those who are away from home and, hence, omitted from the survey. There are several techniques for dealing with this difficulty, all aimed at an estimate of how the non-respondents would have responded had they been reached and interviewed.

As has been pointed out, however, no sample is ever without shortcomings: the exigencies of costs, accidents, and other circumstances may escape control and introduce bias. It is the expert's preeminent task to enlighten the court as to the relevance of such flaws in respect to the measured issues. The point is an important one, for even though a sample have many flaws, it may be judged sufficient for deciding a particular issue. The surveys of the late Dr. Kinsey are illustrative although, of course, they have not come before the courts. The "Human Male" and "Female" were represented only by those ill-assorted men and women who, by one means or another, could be persuaded to be interviewed.

There is a standard method of estimating the survey's true value from such an improperly selected sample. First, the degree of under- or over-representation of certain sub-groups is determined, *e.g.*, too many college educated, not enough laborers. One then estimates what the over-all, corrected group average *would* have been had the sub-groups been represented in their true, known proportions.[61] While, of course, such estimates lack the precision of probability samples, they will often be satisfactory if the group measure clearly falls beyond the crucial minimum requirement. But more often, if a sample is improperly designed, the expert will be unable to appraise the size of its bias.

---

[60] It is bad practice to permit the survey organization to present its best interviewer to the court; a survey is only as good as its weakest link.

However, even a properly designed and well executed sample may prove wanting, simply because it is too small to provide the desired response, *i.e.*, the sampling error may be too great. This difficulty may derive from the fact that a sample can be sufficiently large to answer some questions, but too small to answer others. Here, again, only the expert can advise the court with precision.[62]

Circumstances of the Interview

The third point at which the validity of a survey must be tested is at its line of questioning and the circumstances under which the interview was conducted. Lawyers know that there is more than one way of posing a question and that the response in each case may be different. Such differences may result from the phrasing of the individual questions, from their sequence, or from the questioning situation. There is a body of experience from which the expert will be able to guide the trier of facts. He will detect bias where the layman sees none, he will know where memory failure will tend to underrate and where vanity may have the opposite effect, and he will know also when, at times, the interviewer's personal opinion affects his respondent's answers.[63]

Yet it is axiomatic, in survey technique, that the danger of question bias increases with the complexity and ambiguity of the questions. If their aim is simple and factual, such as determining the make of the respondent's automobile, neither the form nor the sequence of the questions will make much difference. But in the survey question "As a guess, how much wax would you say there was in glass wax . . . ," the court rightly found "a built-in bias."[64]

---

[61] An example of this procedure can be found in Zeisel, "Sexual Behavior in the Human Female," 21 U. Chi. L. Rev. 517, 519 (1954).

[62] It might be remembered that the issue of the adequate sample size becomes relevant only with a good sample. As the Literary Digest experience showed, if a sample is wrongly designed its being large makes it only worse, because it makes it less probable that the systematic bias in sampling is cancelled, or at least reduced, by the sampling error working accidentally in the opposite direction.

[63] Hence, the requirement that the interviewer know as little about the purpose of the survey as is compatible with his duties. For a complete examination of all such interviewing problems, compare Hyman, Interviewing in Social Research (1954).

In another case it was the *sequence* of questions which was found to color the results. At issue was the confusion of "All State" and "All States Life Insurance":

(3) What does 'All State' mean to you?
(4) If you wanted All State Insurance where would you go?
(5) Have you ever heard of All States Life Insurance Company?
(6) Who would you say owns All States Life Insurance Company?

The court, with justification, criticized the survey for "not fairly presenting the name All State."[65]

Other aspects of an interview can also become grounds for criticism. Word association tests given to students in a classroom were rejected because their reactions were "bound to differ from that of the buyer in the market place when confronted with the . . . beverage. . . ."[66] As another court remarked, "the issue is not whether the goods would be confused by a casual observer, but [rather] . . . by a prospective purchaser at the time he considered making the purchase. If the interviewee is not in a buying mood but is just in a friendly mood answering a pollster, his degree of attention is quite different."[67]

A rather subtle source of bias was noted by a court[68] in the selection, as the interviewing area, of "the vicinity of Syracuse, which is not far . . . from the town in which plaintiff's goods are manufactured," thus producing an abnormally high degree of confusion with his brand.

Occasionally, the problem at issue will present genuine difficulties to the interviewee. Such an issue arose at a recent hearing before an FCC examiner concerning the sale of radio station WGMS in Washington, D.C. When survey results purporting to show the audience's preference for "classical" and "semi-classical music" were presented, it was contended that the respondents did not understand the meaning of these terms.[69] The problem of ambiguity might be avoided by posing a *battery* of questions, each of which would cover one facet of the ambiguous concept.[70] Together they would insure that the interviewee has responded to all aspects of the concept.

---

[64] S. C. Johnson & Son, Inc. v. Gold Seal Co., 230 F.2d 832 (D.C. Cir. 1956), 40 Trademark Rep. 347, 357-58 (1950), aff'd, 90 U.S.P.Q. 373 (1951).

[65] Sears Roebuck & Co. v. All States Life Ins. Co., 246 F.2d 161, 171, 172 (5th Cir. 1957).

[66] Coca-Cola v. Nehi Corp., 27 Del. Ch. 318, 326, 36 A.2d 156 (1944).

[67] Judge Wyzanski in American Luggage Works v. United States Trunk Co., supra note 37 at 53.

[68] Oneida, Ltd. v. National Silver Co., supra note 46 at 286.

[69] 52 Broadcast Telecasting Magazine 68, February 11, 1957.

[70] Compare Lazarsfeld, Comment appended to the Blum and Kalven article, supra note 3.

The hearsay problem also may arise as a result of survey techniques. In *Orbo Theatre Corp. v. Loew's, Inc.*,[71] the trial court refused to admit a survey in which the respondents were asked about their movie going habits *and* those of their families. This added level of hearsay constituted one of the barriers for admittance of the survey.

By way of conclusion, then, we may say that a survey can err in at least three basic ways:

(1) it may aim at an irrelevant universe;
(2) although aiming at the right universe, it may not be representative because of faulty sampling, or it may be based on too small a sample and hence render measurements that are not sufficiently precise;
(3) the mode of questioning, the interviewing situation, or the sequence of questions may tend to reflect inaccurately the characteristics at which the survey aims.

EXPERTS AND PROCEDURAL SAFEGUARDS

Courts and juries are presumed capable of judging the value of witness testimony. They are trusted to determine if such testimony is irrelevant or otherwise defective and without weight. Because of a tendency to present survey findings in a way that makes them appear simple and judgeable by the layman, the detection of defects in survey evidence and the evaluation of its weight and significance is not a simple task. Hence, no survey should be presented without accompanying expert testimony. The expert will be needed precisely because of the surface simplicity which almost all surveys, good or bad, tend to display. Thought should also be given to the possibility of appointing impartial survey experts, either by agreement of the litigants or by choice of the court. They would be analogous to impartial medical experts.[72]

Survey evidence has still another peculiarity which should guide its use in court: its production is usually very costly and time-consuming. To confront the adversary with a survey only at the time of trial will almost always constitute an unfair surprise. Surveys, in this respect, ought to be treated much like experiments which, in purpose and mode of analysis, they closely resemble.[73] Two procedural suggestions offered by Professor McCormick in regard to experimental evidence can apply without modification to survey evidence:

---

[71] 156 F. Supp. 770 (D.D.C. 1957).

[72] See Chapter 11, discussing Impartial Medical Experts in Zeisel, Kalven and Buchholz, Delay in the Court, (1959); see also Zeisel, "The New York Expert Testimony Project: Some Reflections on Legal Experiments," 8 Stan. L. Rev. 730 (1956).

[73] McCormick, Evidence § 169 (1954); Zeisel, Say It With Figures 132 (4th ed. 1957).

> . . . [T]he adversary system . . . must be modified . . . by a rule of court providing first, that no experiment [survey] shall be used in evidence unless reasonable *notice* shall have been given the adversary, with an opportunity to make suggestions as to planning and to be present at the test; and second, empowering the court, in discretion, on application of either party, to appoint an *impartial* person to conduct . . . an experiment [survey].[74]

Where both sides can agree on an impartially conducted survey, this would have the added advantage of dividing the survey costs.[75]

TOWARD A SURVEY EXCEPTION

The law with respect to survey evidence is still far from settled doctrine. Thus far, the development has been guided by fears that, since most surveys are hearsay evidence, a bad one might too easily mislead the trier of facts. But our discussion has shown that while the dangers of an uncritically received survey are real enough, they derive not from its hearsay character, but primarily from elements easily opened to expert review. If such expert help is available to the court and the parties to the trial, the dangers arising from the admittance of survey evidence are much smaller than is reflected by the rules which presently govern their admission. These dangers will become negligible if, in the preparation and presentation of survey evidence, the following safeguards are observed:

(1) All sampling plans, instructions to field workers, questionnaires and other survey instruments ought to be available as evidence of its design.

(2) The survey staff, from the director down to the ultimate field workers, should be available for questioning as to the survey's manner of execution. The survey interviewees, as a rule, ought not to be required to testify.

(3) The survey evidence should be presented by an expert witness.

(4) If a survey is planned during the course of the litigation, the court should explore the possibility of having the survey conducted by stipulation of parties through an agreed-upon or court appointed impartial expert. At that time, such technical requirements as size of sample and other specifications could also be stipulated. If this should not prove feasible, a litigant intending

---

[74] McCormick, Evidence § 169, p. 362 (1954).

[75] . . . [W]here contending parties . . . can agree on the making of a survey (which may or may not mean agreeing to be bound by the *results* of the survey), many if not most of the really controversial problems about survey-making will be eliminated.

Waterbury, "Opinion Surveys in Civil Litigation," 44 Trademark Rep. 343, 361 (1954). See also Note, 66 Harv. L. Rev. 498 (1953).

to offer a survey in evidence should be required to notify his opponent early enough to enable him to become an observer in its development. If the survey was completed *prior* to the commencement of the litigation, it should be disclosed to the adversary well in advance of the trial.

If these safeguards are provided, the court should be satisfied that the evidential value of the survey can be appraised objectively; hence, nothing should prevent its admission, provided it is relevant to the litigated issue. If these safeguards are not provided, the court ought to refuse to admit the survey.

While the law might ultimately develop in this direction, its present state gives only small encouragement. A reading of the Model Code of Evidence and the proposed Uniform Rules of Evidence shows quite clearly that the isolated cases in which survey evidence was admitted are far from developing into a settled doctrine. The Uniform Rules do not admit surveys unless they qualify under the state-of-mind exception or, perhaps, as "entries made in the ordinary course of a business," this exception would, at best, apply only to surveys made prior to, and unconnected with, the litigation. However, since surveys provide the best, if not the only, evidence on certain issues, and since expert knowledge in the field has advanced sufficiently to protect the trier of the facts from error, the law may well lower its heavy guard.

UNIVERSE AND SAMPLE I I

The necessary compromise enforced by limited resources which Mueller and Schuessler refer to in their discussion of sampling procedures is clearly illustrated in the following reading. By defining his universe as "all individuals injured or killed in auto accidents during a calendar year," Conard was forced to use sampling techniques. However, as he points out, "The ideal sampling source . . . did not exist; however, three potential sampling sources were available. These were insurance company records, hospital reports, and police accident reports." And, as a consequence, the sample drawn would necessarily be biased. Insurance company records do not include injured non-insured persons, hospital records exclude individuals treated in doctors' offices, and not all personal injury accidents are reported to the police.

In addition to this problem of sampling sources, there is another aspect of the reading which is useful to the student of social science methodology. The author takes great pains to make it clear how he drew his samples and collected his data. (In fact, his methodological narrative is a good example of what Howard Becker called a "natural history.") Conard presents the nature of his procedures at each stage of the research, and any challenge against his findings can be directed at the methodological structure of his study instead of trailing off into a vague skepticism aimed at the social sciences in general.

From Alfred F. Conard, *et al.*, *Automobile Accident Costs and Payments* (Ann Arbor: The University of Michigan Press, 1964) pp. 9-11; 322-366.

Alfred F. Conard, *et al.*, *Automobile Accident Costs and Payments*
(Ann Arbor: The University of Michigan Press, 1964)
pp. 9-11; 322-381

The study was born of a suspicion that the handling of personal injury cases is among the most critical problems facing the legal profession today. Clearly a large part of the public believes that ambulance-chasing and outrageous fees are commonplace. The waiting period to get to trial in many major cities is notorious. Even legal theory is showing symptoms of malaise, as attacks are made on contributory negligence, damages for pain and suffering, and exclusion of evidence of insurance.

The designers of this study did not wish to contribute to the welter of opinions on legal theory, nor even to add to the studies of client procurement, attorney compensation, and trial delays. They conceived the notion that a new start should be made by studying the underlying human demand whose pressures have bubbled forth in the form of a "fee problem," a "delay problem," and other "problems." This human demand was conceived to be the desire for something to fill the trench in material well-being which is gouged by a personal injury.

The grand design was to discover what are the economic losses from injury, and what is being done to repair these losses. It was supposed that the trail would lead back to the point of initial curiosity—injury litigation; but it might lead in a good many other directions, which would be equally instructive.

Three separate methods of study were adopted. The first method was to collect and analyze national statistics on programs which would presumably come to the aid of an injury victim. The results of this approach constitute Part I of this report.

The second method was a field survey. It began with interviews with persons involved in personal injury automobile accidents. Later, thanks to an additional grant of funds, it was extended to include interrogation of lawyers for the injury victims, lawyers for defendants, individual defendants, and hospitals. This led to the heart of the study, which is the survey of Michigan automobile injuries, reported in Part II.

The third method was an inquiry into foreign systems for dealing with the same human demand. Foreign laws, lawyers, courts, and insurance companies may differ from their American counterparts, but modern foreign countries are sure to have the same human demands, occasioned by accidental injuries. Informants from England, France, Sweden, and West Germany supplied information on sources of reparation for automobile injury victims in their respective countries.

Alfred F. Conard

Survey Methods: A Description

INTRODUCTION

This chapter describes the research design developed for the Michigan survey. Each of the independent components of the final design is presented separately, with emphasis on the underlying design considerations, sample composition, questionnaire content, and response and weighting characteristics for each. These discussions are preceded by a description and definition of the group of individuals (universe) being studied, as well as an overview of the total survey design. The present chapter also describes the "weighting" procedures used for the various parts of the study, the procedure used to substitute plaintiffs from the court sample for plaintiffs from the police sample, and the procedure used to estimate income loss for sampled individuals, and the techniques used to collect and process the data.

The following chapter evaluates the research design described here, placing special emphasis on the extent to which the assumptions and design specifications could have introduced serious inaccuracies into the final results.

A. THE UNIVERSE DEFINED

The universe (the total group of persons being studied) included all individuals injured or killed in automobile accidents that occurred in the State of Michigan during one calendar year, as well as noninjured drivers and owners involved in these accidents. The latter were included so that the estimate of the total economic loss resulting from personal-injury automobile accidents would include an accurate valuation of property damage which might have been compensated along with the personal injuries. To state this another way, the sample for this study was selected so that inferences could be made about various characteristics of all personal-injury accidents occurring in Michigan during one calendar year. A calendar year, or some multiple thereof, was desirable in view of the fact that available Michigan and national traffic accident and court statistics are collected and published on a calendar year basis. The final sample is a statewide area probability sample which is representative of the universe of individuals as defined above.

B. Survey Design: An Overview

The survey design for this study includes two principal samples, one from police files and the other from court records. The first, which provides the basic sampling frame for the entire study, is a probability sample of automobile accidents which occurred in Michigan during 1958, and which were reported to either the Detroit Police Department or the Michigan State Police as having resulted in injury or death to one or more persons. This sample consists of 1118 accidents, which resulted in 2872 individuals being listed on police accident reports. The unit of analysis for most of the data presented in the report is the injured individual, not the accident.

Summary data concerning economic losses, injuries, and legal actions taken by these individuals were collected using a mail questionnaire, with a mail and telephone follow-up for those not responding.

The returns were grouped into three categories. The first category included individuals for whom the completed questionnaire provided adequate data for analysis; i.e., in terms of the purposes set forth for the study, it was felt that no additional data would be required from these individuals. Persons in this first category had medical expenses of less than $500 and incurred no permanent physical disability or permanent impairment of ability to work.

The second category included individuals who either died as a result of the accident or who sustained complicated or "serious" personal injuries—usually involving large medical expense, a permanent physical disability or both. For most of these individuals and their families, the accident marked a major turning point in their lives. Most of these individuals required extensive medical treatment, often followed by job retraining or rehabilitation. Serious financial difficulties were frequently encountered. For such persons, information provided by the mail questionnaire was not sufficiently detailed to permit a complete analysis of the accident and its consequences. Accordingly, these individuals were designated to be reinterviewed with a more detailed personal-interview schedule.

The third category of mail questionnaires consisted of those individuals who had retained a lawyer and who were involved in some kind of legal action at the time the questionnaire was completed. Since, at the time the respondents were asked to complete the mail questionnaires, the accidents being asked about were, on the average, only two years old, many of the legal actions were just being initiated, and very few had been completed. Persons involved in these legal actions would not be in a position to give complete cost and compensation data concerning their accidents until the cases had been settled. In fact, most of these persons had been instructed by their lawyers or insurance companies not to discuss the case with anyone. Since many of the cases would remain in litigation for another two or three years, it was decided to drop this group entirely and replace it with an independent sample of older automobile personal-injury cases, the majority of which would be settled.

In order to fill the gap created by this elimination and also to secure a broader sample of litigated cases, the second of the two principal samples for this study was used. It was a probability sample of automobile personal-injury suits filed in 1957 on the calendar of either a Michigan Circuit Court, the Kent County Superior Court, or a Federal District Court located in Michigan. Plaintiffs in the sampled cases were used to represent individuals in the original police sample who were plaintiffs in a court action. The exact procedure used to substitute one sample for another is described in detail on pages 345-50 of this chapter, and the results of the substitution are evaluated in Chapter 10.

The plaintiffs in the court suits, the respondents returning mail questionnaires for whom additional data were desired, and a sample of those who had not responded to the mail questionnaire were combined into one personal-interview sample. Personal interviews were completed with 406 of the 564 designated respondents; 28 of the 406 were victims of nonserious accidents who had not returned the mail questionnaire.

The combined data provided by the questionnaires outlined above were considered sufficient to meet the purposes originally outlined for the study. However, in the course of processing the interviews, it was discovered that respondents were often unable to provide certain types of information asked for on the questionnaires. For example, in cases where insurance companies had paid hospital and medical bills directly, the injured individual or his family were frequently unaware of the amount paid. In other cases, the individual being asked about had been killed in the accident or had died since the accident, and the respondent was some other member of the family. Such respondents possessed varying degrees of knowledge concerning the facts of the accident; in some cases they knew nothing about injury costs incurred and compensation received.

After careful consideration of apparent gaps in the data, it was deemed desirable to secure certain types of information—particularly financial data—from other participants in the compensation process. Such data would provide missing information in some cases and verification of respondents' reports in other cases. Accordingly, a second research grant was secured for these purposes. The additional field studies undertaken are outlined briefly below in chronological order.

The first study included personal interviews with two groups of claimants' lawyers. The first group included all lawyers shown on the sampled court records as representing plaintiffs, whether or not a personal interview had been obtained with the plaintiff. The second group included lawyers hired by those individuals sampled from police records who had both (1) completed a personal interview and (2) granted permission for their lawyer to be interviewed. There were 63 lawyers in this second group. Claimants' lawyers were questioned about legal proceedings and strategy, as well as about costs and compensation to both themselves and their clients. A few of the claimants' lawyers answered more than one questionnaire, because they represented more than one sampled plaintiff.

A second set of questionnaires was completed by telephone with the individual defendants shown on each sampled court calendar. The purpose of these interviews was not to determine the total effect of the accident on the defendant and his family, but only to examine the process by which the defendant undertook to defend himself and the direct consequences of the process itself. Each defendant was asked about his involvement in legal proceedings (including the manner in which he secured counsel) and about any psychic or economic losses incurred by himself or his family as a result of the suit.

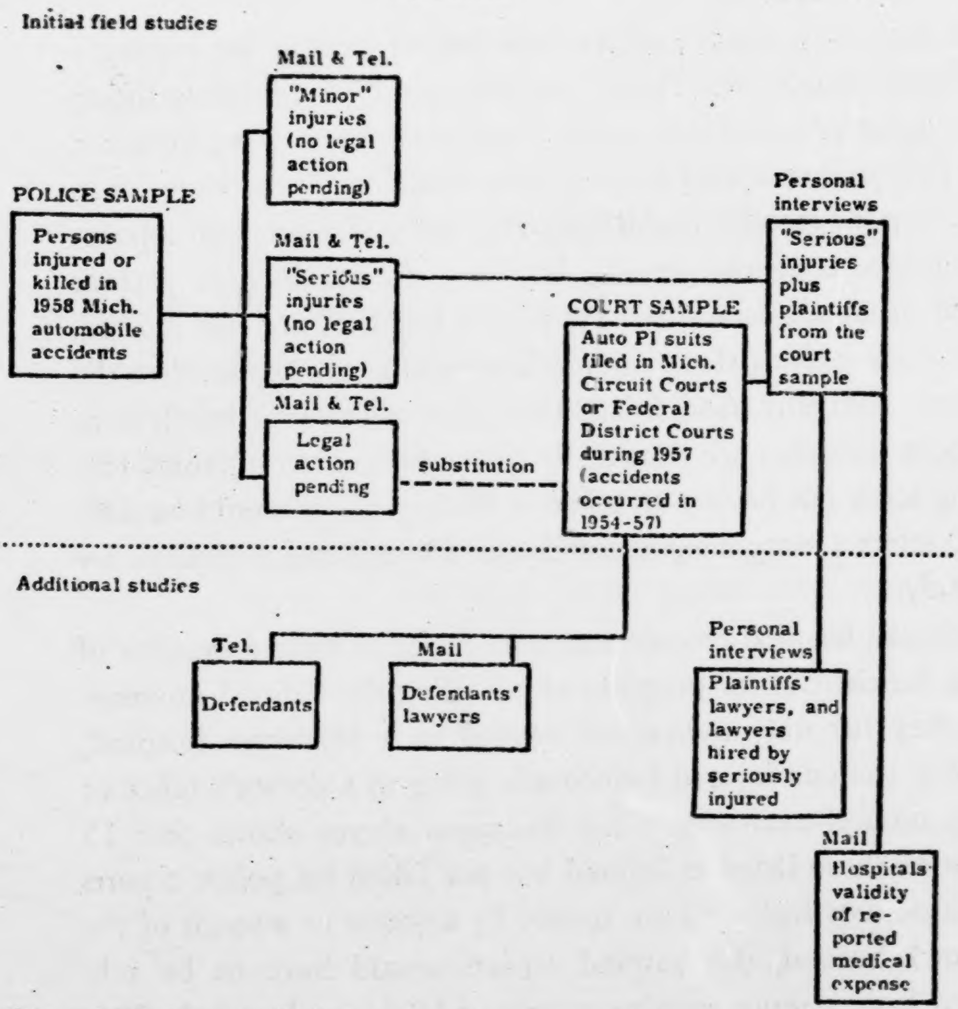
A third study involved mailing a two-page questionnaire to each lawyer listed on the sampled court records as representing a defendant. Questions asked were parallel to questions already asked of plaintiffs' lawyers, but much fewer in number. They were primarily concerned with the major issues or sources of disagreement in the case and the important factors underlying determination of a final settlement. As among the plaintiffs, but to a much greater degree, some lawyers or law firms represented several of the defendants in the sample.

A final set of mail questionnaires was sent to all hospitals, clinics, or other medical institutions named by personal-interview respondents as having treated or cared for the injured or deceased individuals. Hospitals were asked to complete a separate one-page questionnaire for each discharge subsequent to the date of the accident. Questionnaires asked about the total hospital bill for each visit, how much of the bill had been paid, and who paid it. These data have been used mainly to examine reporting bias.

Figure 9-1 summarizes the above overview. In the immediately following sections of this chapter, sample design, questionnaire content, and field results for each part of the study will be dis-

FIGURE 9-1

Study Design*


Initial field studies
Mail & Tel.
"Minor" injuries (no legal action pending)
POLICE SAMPLE
Persons injured or killed in 1958 Mich. automobile accidents
Mail & Tel.
"Serious" injuries (no legal action pending)
Mail & Tel.
Legal action pending
substitution
COURT SAMPLE
Auto PI suits filed in Mich. Circuit Courts or Federal District Courts during 1957 (accidents occurred in 1954-57)
Personal interviews
"Serious" injuries plus plaintiffs from the court sample
Additional studies
Tel.
Defendants
Mail
Defendants' lawyers
Personal interviews
Plaintiffs' lawyers, and lawyers hired by seriously injured
Mail
Hospitals validity of reported medical expense

*Non-response groups omitted

cussed in more detail, followed by a description of the procedures used to combine the various parts of the study for analysis purposes.

C. THE POLICE SAMPLE

1. *Sample Design Considerations*

As previously indicated, the universe for the total study included all individuals injured or killed in automobile accidents that occurred in the State of Michigan during one calendar year. Before describing the study design actually used, a number of the more important problems encountered in making this definition operational will be outlined.

An important initial problem was that of locating an appropriate sample source. The "ideal" sample source—a complete listing of Michigan personal-injury auto victims—did not exist; however, three potential sampling sources were available. These were insurance company records, hospital reports, and police accident reports.

Insurance company records, however, do not include persons injured in an accident when the parties involved are not insured, nor do they include those injured in accidents not reported to the insurance company. And, in addition, since individuals involved in Michigan accidents are insured by many different companies (including some not having an office in Michigan), it would be difficult to select a sample representative of the universe as defined for this study.

Michigan hospital records also were found to have a number of serious deficiencies for purposes of sampling the defined universe. First, they list only individuals treated in a Michigan hospital; thus they exclude injured individuals going to a doctor's office or relying on self-treatment. (The Michigan survey shows that 15 percent of those listed as injured but not killed on police reports stated that they had not been treated by a doctor as a result of the accident.) Second, the hospital reports would have to be subsampled from a representative sample of Michigan hospitals. This would involve negotiating for access to files with each sampled hospital on an individual basis. Policy regarding admission to files varies from hospital to hospital, as do the administrative procedures. The sampling procedures within any one hospital would have to be individually tailored to the record-system used.

Another crucial sampling problem within each hospital would have been that of identifying injuries resulting directly from automobile accidents. Such identification would have been based on the hospital's admitting diagnosis which might or might not make a direct reference to the automobile accident causing the injury. (It might be expected that lack of direct reference would be particularly prevalent in cases where the injury did not manifest itself for some time after the accident, or in cases where numerous hospitalizations were required for recovery.) It is conceivable, if not inevitable, that serious error would result from the actual mechanics of reading and classifying entries on the admission papers. And finally, in terms of the sampling the defined universe, error would be introduced by the fact that an individual injured in a Michigan accident could be hospitalized in another state, and an individual injured in another state could be hospitalized in Michigan.

For purposes of drawing a representative sample of individuals injured in personal-injury automobile accidents, police accident reports were considered deficient in only two minor respects. First, not all personal-injury accidents are reported to the police. Such non-reporting often occurs either because no police official is immediately available, or because the parties involved mutually agree to settle their differences, often to avoid a ticket. Second, many individuals do not become aware that they have been injured until after the police report has been completed and all parties have left the scene of the accident. This is often the case with minor back or internal injuries. Even though many of these accidents are reported to the police (as involving property damage), the individuals do not appear on the police reports as having been injured.

An evaluation of the above and other considerations indicated that the best available sampling source would be police accident reports. The major reasons for this decision are as follows. First, the actual selection of the sample could be accomplished by securing access to only two sets of records. Personal-injury and fatal accident reports for the City of Detroit are filed at Detroit Police Headquarters. Reports for the rest of the state are filed at State Police Headquarters in East Lansing. Second, the police reports offered the only sample source consistent with the defined universe. Third, the police accident reports are, by law, a matter of public record. (Accompanying documents, such as signed statements or police investigation reports, are not available to the public.) A discussion of the possible biases present in such a sample will be found in Chapter 10; it should be indicated here, however, that the biases have been evaluated as relatively small.

Before the sampling of police reports could be begun in Detroit and East Lansing, there remained the problem of which particular calendar year (or years) to study. Preliminary studies had indicated that a high percentage of all Michigan personal-injury automobile accidents reported to the police could not be considered "serious" when measured by any yardstick (e.g., number of days in the hospital, time lost from work, property damage, compensation received, etc.). The large majority of all personal-injury accidents appeared to be settled in a relatively few months after the accident. For "minor" accidents, it was felt that a long time span between the date of the accident and the date of interviewing would invite the possibility of serious error from memory distortion; it seemed likely that respondents would not be able to report accurately an event which had occured a number of years in the past, and which was not important to them. Indeed, there also would be serious problems in locating individuals with very old address records.

The preliminary studies showed that a large fraction of the economic loss incurred in personal-injury automobile accidents resulted from the relatively small number of "serious" accidents, many of which (particularly those going to court) required up to six or seven years to settle. For the purposes of this research, it was essential that most of the cases selected be settled at the time of the interview.

2. *Final Sample Design*

The final sample of police personal-injury reports is a statewide representative sample of accidents that took place during 1958. The actual sampling of police files resulted in the selection of 1118 accidents, which included 2872 individuals eligible for the study.

The basic *unit of analysis* around which the final research design was constructed is any individual listed on the sampled accident reports as having been injured or killed in an automobile accident that took place in Michigan during the 1958 calendar year.

The sample of police reports was drawn from two sources. The Detroit Police Department provided records for nonfatal personal-injury accidents that took place within the city limits, and the Safety and Traffic Bureau of the Michigan State Police provided records for nonfatal accidents taking place outside of Detroit and for all fatal accidents in the state. Nonfatal personal-injury accidents were selected using a sampling rate of one in forty-two; fatal accidents were sampled at a rate of one in six.

These police reports constitute the basic sample frame for this study. However, as indicated previously, one subset of eligible respondents listed on these reports was dropped for purposes of analysis, and a substitution was made using a sample selected from the Michigan courts. The court sample is discussed in detail in a subsequent section of this chapter.

D. MAIL AND TELEPHONE SCREENING INTERVIEWS

1. *Purpose*

Only a few personal-injury automobile accidents are serious. For most minor accidents, answers to a few factual questions will provide a relatively complete picture of the accident and its related costs. To avoid the high cost of conducting personal interviews with a large number of accident victims who had only minor injuries and who received little or no reparation, a screening questionnaire was mailed to all individuals sampled from police records. The questionnaire was designed to determine basic factual data concerning the extent of both the injury and the economic loss resulting from the accident. Individuals who did not return the first form were sent a second mail questionnaire, and those who still did not respond were telephoned by members of the Survey Research Center's field staff. For an analysis of the reliability of the screening questionnaire, see Chapter 10.

2. *Questionnaire Content and Eligible Respondents*

The mail screening questionnaire accompanied by a cover letter was sent to each eligible respondent believed to be living at the time of the mailing, and to the next-of-kin of individuals listed as fatalities on the police reports. In general, questionnaires asking about deceased individuals were sent to the "Informant" listed on the Death Certificate. (In Michigan, a photostat of the Death Certificate for each listed fatality is filed with the police accident report.) An accompanying letter explained the purpose of both the study and of selected questions on the questionnaire. It guaranteed the respondent anonymity and it instructed the respondent not to answer certain questions if a lawyer was still working to help collect money in connection with the accident. The questionnaire itself asked about medical treatment and medical expense, legal aid received and legal expense, the number of days lost from work, valuation of property damage, whether or not any permanent disability had been incurred, and the sources and amounts of reparation received. In the case of fatalities, additional questions were asked about the age, education, and income of the deceased individual at the time of the accident.[1]

The content of the telephone questionnaire was identical to that of the mail questionnaire; only the questionnaire format was altered to facilitate asking questions and recording answers by phone. If the injured individual was under sixteen years of age at the time of the interview, the interviewer was instructed to interview a parent or any other responsible adult with knowledge of the accident.

3. *Results*

The results of the initial screening are shown in Table 9-1. Of the 818 individuals by whom no questionnaire was completed 63 percent were never located, despite an extensive search of telephone books and city directories.

The screening questionnaire was not designed to provide complete information about seriously injured individuals. Further information was to be secured from them by personal interview.

TABLE 9-1

Sample Size and Response Rate for Mail and Telephone Screening Questionnaire

	Number of questionnaires
Number of questionnaires completed	*2054*
By mail	1287
By telephone	767
Number of questionnaires not completed	*818*
Total sample	2872
Percent of questionnaires completed	71.5%

---

[1] Space prohibits the inclusion of most of the actual documents used in this research. However, copies of the cover letters, questionnaires, interviewers' instructions, editing worksheets, and codes for each part of the study have been microfilmed, and copies of the film can be obtained from the Librarian, Institute for Social Research, The University of Michigan, Ann Arbor, Michigan. (Reference: Supplement A, Michigan Automobile Study.)

For purposes of this study, a "serious" injury was defined as one which resulted in death, or resulted in medical bills of $500 or more, or required hospitalization of three weeks or more, or resulted in some permanent impairment of ability to work. Two hundred ninety-four of the respondents who completed a screening questionnaire were classified as seriously injured and subsequently were designated for personal interviews. For the "minor" injuries, the data provided by completed mail or telephone questionnaires were considered to be sufficient for analysis. Additional personal-interview respondents were forthcoming from the Michigan court sample.

E. THE MICHIGAN COURT SAMPLE

1. *Sample Design*

Because of the recentness of the accidents, the police sample included a group of individuals who were still involved in some form of legal action at the time of the initial mail and telephone interviewing. These individuals might be expected to be either reluctant or unable to give complete cost and compensation data concerning the accident until legal actions were concluded. Rather than wait for completion of these proceedings, it was decided to draw a second sample of older personal-injury automobile accidents which involved court action. How was this substitute sample selected?

Initially, it was established that calendar year 1957 was likely to be the most recent year for which most cases commenced in that year could be traced to their final disposition. Therefore the universe for the court sample was defined as all personal-injury suits filed during 1957 on the calendar of either a Michigan Circuit Court, the Kent County Superior Court, or a Federal District Court located in Michigan.

In order to sample this universe, a number of important practical problems had to be overcome. First, there are no statistics available in Michigan which describe the universe as defined.

Statistics describing the business of the Michigan Circuit Courts are collected and published by the Administrator of State Courts. These statistics are broken down only among law, chancery, and criminal calendars. There are no figures available which describe the composition of civil litigation by type of case. In order to determine the proper sampling fractions for the Michigan court sample, a separate study, the Michigan Court Study,[2] was undertaken to examine the law and chancery calendars of Michigan Circuit Courts.

To insure efficient use of available funds, time, and research personnel, stratified probability sampling techniques were used (1) to select twenty-three counties in the state and (2) to subselect 2411 cases from within these counties in such a manner that the cases included in the sample would be an unbiased representative sample of all cases filed in the state. Independent samples were drawn from the law and chancery calendars. In order that the final sample include about the same number of cases from each calendar, it was decided to sample one out of each sixteen cases filed on civil law calendars and one out of each twenty-eight cases filed on chancery calendars. Data from the court calendars were transcribed during February and March of 1960.

The final court sample included 1226 cases from the chancery calendar and 1185 cases from the civil law calendar. Of the civil law cases, 256 were filed to recover damages resulting from a personal-injury automobile accident. Since the court action for all but 11 of these cases had been completed at the time the sample was drawn, and since 256 cases were more than enough for the desired auto personal-injury sample, these cases were designated as the substitute sample.

To insure adequate representation of cases involving out-of-state plaintiffs or the federal government, an additional sample was drawn from the three Federal District Court Divisions located in Michigan. The calendars of these courts indicate the cause for the complaint; consequently, the sample was taken from cases indicated on the calendar as personal-injury automobile accident cases. No winnowing of other kinds of cases was necessary. Selected at a rate of one in four automobile personal-injury cases, the final sample included 53 suits, of which three were still open at the time the sample was selected.

---

[2] For a brief discussion of the results of this study, see Alfred F. Conard and Charles E. Voltz, "The Economics of Injury Litigation," Michigan State Bar Journal (August 1960).

Thus, the composite court sample was made up of 309 cases—256 from the Circuit Courts and the Kent County Superior Court, and 53 from the Federal District Court Divisions. However, the budget would not allow inclusion of all these cases in the personal-interview sample, and a subsampling procedure was required. The cases were divided into two groups. The first included cases where the plaintiff filed for damages in excess of $25,000. This group consisted of 123 cases and was included in the personal-interview sample with certainty, i. e., all 123 were designated for personal interviews. The second group included all cases in which the plaintiff filed for damages of $25,000 or less. The 172 cases in this group were subsampled at a rate of one in two, with every other case (N = 84) being included in the personal-interview sample. The 14 "still open" cases were excluded from the personal-interview subsample, since it would not be proper to interview the persons concerned before a final settlement had been reached. The 14 cases are not represented in the analysis. The final personal-interview sample included 207 court cases. Table 9-2 summarizes the composition of the court sample. To the extent that the 14 "still open" cases are atypical, a bias is introduced. A follow-up could be completed on these cases at some future date.

The reader should be aware of one important methodological problem associated with combining data from the police and court samples. Michigan police reports include only accidents

TABLE 9-2
Composition of the Michigan Court Sample

Court	*Time required to settle*	*Size of suit*	Number of cases *Total*	Number of cases *In personal-interview subsample*	Number of cases *Not in personal interview subsample*
Circuit and Kent Superior	2 years or more	More than $25,000	55	55	—
		$25,000 or less	63	31	32
	Less than 2 years	More than $25,000	43	43	—
		$25,000 or less	84	40	44
Federal District	2 years or more	More than $25,000	13	13	—
		$25,000 or less	11	5	6
	Less than 2 years	More than $25,000	12	12	—
		$25,000 or less	14	8	6
Case still open			14	—	14
Total number of cases			309	207	102

that occur within the geographic limits of Michigan. However, some Michigan accidents involve vehicles from other states; and it is possible that a Michigan resident would choose to sue an out-of-state motorist in the state of the latter's residence. On the other hand, any sample of automobile accident cases filed in Michigan courts will include some suits resulting from accidents which did not take place in Michigan. However, in this study both of the groups described are small relative to the total sample, and it is highly unlikely that any extensive biases are introduced by making the assumption that court cases filed in Michigan resulting from accidents occurring outside the state are equivalent to out-of-state court actions arising from Michigan accidents. For purposes of the analysis herein, this assumption has been made.

The procedures used to substitute plaintiffs from the court sample for plaintiffs from the police sample are described later in this chapter, and the effectiveness of the procedures is evaluated in Chapter 10.

F. PERSONAL INTERVIEWS WITH PLAINTIFFS AND PERSONAL INTERVIEWS WITH MAIL AND TELEPHONE RESPONDENTS REPORTING "SERIOUS" INJURIES

1. *Purpose*

Personal-interview questionnaires were designed to secure information concerning the consequences of serious or complicated automobile accidents, such as those involving death, serious injury, extended litigation, and large income losses.

The 564 respondents designated for personal interviews were selected from three sources. First, there were 292 "serious" cases from the mail and telephone screening interviews. Second, there were 207 plaintiffs from personal-injury automobile suits filed during 1957 on calendars of Michigan Circuit Courts, the Kent County Superior Court, or one of the three Federal District court Divisions located in Michigan. And third, in order to learn something about the characteristics of persons who did not return a mail or telephone questionnaire, a subsample of 66 cases was selected from those not responding to the screening questionnaire. Since extensive effort had been already made to complete mail or telephone questionnaires with this group, it was evident that additional efforts to secure these interviews would represent a relatively inefficient use of resources; consequently, the sample selected was intentionally small.

2. *Questionnaire Content and Eligible Respondents*

In order to interview all eligible respondents, two different personal-interview schedules were required. An "A" questionnaire was designed to be used whenever the injured person, himself, was being interviewed; and a "B" questionnaire was designed for situations where someone other than the injured person was being interviewed, e. g., when the injured person was a minor child or deceased.[3]

For cases sampled from court calendars, interviewers were instructed to talk only with the plaintiff. No substitutions were allowed here. (Of course, the plaintiff may or may not have been the injured person.)

For cases sampled from police records, interviewers were instructed to talk only with the injured person, with the following exceptions. If the injured person had died prior to the interview (regardless of the cause of death), the interviewer was allowed to interview an adult member of the household at the time of the accident who had a reasonable knowledge of the facts. Substitution was also allowed if the injured person was twenty-one years old or younger *at the time of the accident.* Eligible respondents for these cases were classified into two groups. If the person injured was sixteen years old or younger at the time of the accident, the interview was obtained with either a parent or guardian. No other substitution was allowed for these cases. However, if the injured person was between seventeen and twenty-one years

[3] Footnote omitted.

old (inclusive) at the time of the accident, the eligible respondent could be either the injured person or a parent or guardian, or the injured person's spouse.

In cases where individuals other than the injured person qualified as eligible respondents, every effort was made to complete the interview with the eligible respondent most knowledgeable about the facts of the accident.

3. *Results*

A summary of sample sizes and response rates for each of the personal-interview groups is presented in Table 9-3.

As shown in Table 9-3, personal interviews were completed with 33 of the 66 individuals in the nonresponse subsample. Two alternatives were available for using these completed interviews in analysis. First, they could be used to represent the entire mail and telephone nonresponse (that is, the 181 cases which did not respond to the screening questionnaire). Second, they could be used as a basis for completing a mail or telephone questionnaire for each respondent, which, in turn, would then be used in the same manner as the original mail and telephone response, that is, to determine if the injury was "serious" enough to qualify the respondent for a personal interview. The small subsample and the large weights that would be involved in the first alternative suggested that a serious bias might be introduced; consequently, the second alternative was chosen. Of the 33 completed interviews, 5 were "serious" and were included with the other "serious" personal interviews for weighting and analysis. The remaining 28 interviews were not "serious" and have not been used in the detailed analysis of personal interviews; but they have been used in conjunction with the mail and telephone returns for the analysis of all cases, thereby improving the response rate. The "adjusted" personal-interview sample, i.e., the one used for analysis, is presented in Table 9-4.

TABLE 9-3

Sample Sizes and Response Rates for Personal Interview Questionnaires

		Interview results		
Sample source	*Number of individuals sampled*	*Number of questionnaires completed*	*Number of questionnaires not completed*	*Percent of questionnaires completed*
"Serious" injuries from police sample	292	247	45	84.1%
Plaintiffs from court sample	207	126	81[a]	60.9
Nonresponse subsample	66	33	33	50.0
Total sample	565	406	159	71.9%

[a] Of the 81 plaintiffs for whom no questionnaire was completed, 62 could not be located.

TABLE 9-4

"Adjusted" Sample Sizes and Response Rates for Personal Interview Questionnaires

		Interview results		
Sample source	*Number of individuals sampled*	*Number of questionnaires completed*	*Number of questionnaires not completed*	*Percent of questionnaires completed*
"Serious" injuries from police sample	297	252	45	84.8%
Plaintiffs from court sample	207	126	81[a]	60.9
Total sample	504	378	126	75.0%

[a] Of the 81 plaintiffs for whom no questionnaire was completed, 62 could not be located.

G. COMPUTATION OF WEIGHTS FOR POLICE AND COURT RECORDS, AND MAIL, TELEPHONE, AND PERSONAL INTERVIEWS

Weighting is necessary when the cases or individuals in the universe have different chances of being drawn into the sample. In this study, for example, the sample of police records consists of every forty-second nonfatal accident record and every sixth fatal accident record. This means that every nonfatal accident selected represents itself and 41 other nonfatal accidents and every fatal accident represents itself and five other fatal accidents. To estimate a statistic for the state as a whole, data from each nonfatal accident record must be multiplied by forty-two and data from each fatal accident record must be multiplied by six. (See column 2, Table 9-5.)

A second weighting step, weighting for nonresponse, may be demonstrated by referring to columns 4 and 5 of Table 9-5. Using the first row as an example, note first that every forty-second nonfatal accident case was sampled. An attempt was then made to complete a mail or telephone interview with each of the 1462 individuals injured in these accidents. *If interviews had been completed with all 1462 people,* the data on each completed interview would have a weight of forty-two. However, only 1075 respondents (73.5 percent) actually completed and returned a questionnaire. These questionnaires were then "weighted up" to represent all individuals injured in Michigan nonfatal accidents during 1958. Thus, the original sample contained 1462 individuals, each representing himself and forty-one other injured persons. After the mail and telephone survey, the same number of injured persons (42 x 1462) was represented by 1075 completed interviews—each respondent representing himself and 57 (42 x 1462 ÷ 1075) other persons injured in nonfatal accidents. The use of identical weighting procedures in all subgroups would implicitly assume that all nonresponse cases were alike and equal to the average of the responses. Such a procedure would be valid if the nonrespondents were similar to the respondents in regard to

TABLE 9-5

Response Rates and Determination of Interview Weights for Information from Police Records and for Mail and Telephone Interviews Completed by Respondents from the Police Sample

Strata	*Initial sampling rate for police records (1)*	*Weight of information from each police record (2)*	*Number of individuals selected for mail or telephone interview (3)*	*Percent of mail and telephone questionnaires completed (4)*	*Weight of each completed mail or telephone questionnaire (5)*
			Personal injury accidents		
Injured	1/42	42	1462	73.5%	57
Noninjured driver	1/42	42	886	73.9	57
			Fatal accidents		
Injured or killed	1/6	6	416	76.4	8
Noninjured driver	1/6	6	108	66.7	9
Total sample			2872	73.8%	

TABLE 9-6

Response Rates and Determination of Interview Weights for Personal Interviews Completed with Respondents from Police Sample

Strata	*Weight of each completed mail or telephone interview (1)*	*Number of individuals selected for personal interview (2)*	*Personal interview response rates (3)*	*Weight of each completed personal interview (adjusted) (4)*
Nonfatal accidents	57	120	88.3%	65
Fatal accidents	8	177	32.5	10
Total sample		297	84.8%	

TABLE 9-7

Response Rates and Determination of Interview Weights for Information from Court Calendars and for Personal Interviews Completed with Plaintiffs from the Court Sample

Strata								
Time required to settle case	*Size of suit*	*Initial sampling rate for court cases (1)*	*Number of court cases sampled (2)*	*Weight for case information from the court calendar (3)*	*Sub-sampling rate (4)*	*Number of plaintiffs selected for personal interview (5)*	*Percent of personal interviews completed (6)*	*Weight of each completed personal interview (7)*
				Circuit or Kent County Superior Court				
2 years or more	More than $25,000	1/16	55	16	1	55	70.9%	23
	$25,000 or less	1/16	63	16	½	31	51.6	62
Less than 2 years	More than $25,000	1/16	43	16	1	43	53.5	30
	$25,000 or less	1/16	84	16	½	40	70.0	46
				Federal District Court				
2 years or more	More than $25,000	¼	13	4	1	13	61.5	6
	$25,000 or less	¼	11	4	½	5	40.0	20
Less than 2 years	More than $25,000	¼	12	4	1	12	66.7	6
	$25,000 or less	¼	14	4	½	8	25.0	32
Case still open			14			0	0.0	
Total sample			309			207	60.9%	

variables important for the study. To the extent that the two groups differ, potential bias exists, and weighting within subgroups can be expected to reduce it. Evidences of bias in this study are discussed in a later section of this book. (See Chapter 10.)

The same reasoning may be applied to the remaining strata in Table 9-5 as well as to the additional tables in this chapter, which show the response rates and weights for the remaining parts of the study.

H. SUBSTITUTION OF PERSONAL INTERVIEWS COMPLETED WITH PLAINTIFFS FROM THE COURT SAMPLE FOR PLAINTIFFS FROM THE POLICE SAMPLE

1. *Procedure for Designating Plaintiffs from the Police Sample*

In order to substitute the personal-interview court sample for plaintiffs from the mail-telephone study (See Figure 9-1, *supra*), it was essential to know which of the 2054 mail or telephone respondents had actually filed a suit. Preliminary drafts of the questionnaire had asked whether a suit had been filed, but experience gained in pretests showed that respondents frequently did not know whether a suit had been filed or not. All they knew was that they had put the matter in the hands of "my lawyer," or "my insurance company." The project staff therefore decided to ask the named lawyer or insurance company for suit information.[4]

It is recognized that an insurance company cannot properly represent a policyholder's personal-injury claim. However, it is probable that the counsel who represents the liability insurer of the injured person, or of the owner of the car in which the injured person was driving would know about any suits filed in connection with the accident. The injured person frequently does not perceive the counsel and the insurance company as two different organizations. This would be especially likely if two car owners are making claims against each other. Because of these considerations, the project staff decided to make inquiry of the insurance company named as "handling" the claim, rather than explaining to the injured person that he cannot be represented by an insurance company and seeking to get from him the name of the lawyer involved.

---

[4] In view of the difficulties encountered in using this procedure, it may well be that in future studies the individuals should be asked directly whether a suit had been filed, and additional information sought only in cases where the individual cannot provide the necessary data.

Before the lawyers or insurance companies were approached, the 2054 completed mail and telephone questionnaires were reviewed to determine which respondents could be reasonably classified as potential plaintiffs. The completed schedules clearly indicated that 1330 respondents had not been plaintiffs. The large majority of this group were eliminated because they answered "No" when asked, "Did you put your case (was the case put) in the hands of a lawyer or insurance company?" The problem, then, was to determine which of the remaining 724 respondents had filed as plaintiffs in a Michigan Circuit or Federal District Court.

If the name of the lawyer or insurance company had been provided by the respondent, a form letter accompanied by a return postcard (shown on the next page) was mailed to the designated party. The letter described the study's purpose, guaranteed the respondent anonymity, and explained both why it was necessary to know the information requested on the postcard and why it was felt that more accurate answers would be obtained from lawyers and/or insurance companies than from the sampled individuals.

Postcards returned by lawyers or insurance companies were sorted into three groups. First, if the completed postcard stated that the individual inquired about had not been a party to litigation of any kind, or had been a party to a suit brought in a lower court (but had not been involved in litigation brought in either a Circuit or Federal District Court), or had been a *defendant* in a suit filed in either a Circuit or Federal District Court, these cards were set aside. For purposes of the sample substitution, it was considered that these persons were not plaintiffs in litiga-

(Please check the *correct* answer) SRC No.

1. Was the individual (named on accompanying letter) either a plaintiff or defendant in a suit filed on a Michigan court calendar? Yes No

IF YES TO #1 2. Plaintiff or defendant? P D

3. In what court was the suit filed?

4. When? Month Year

THANK YOU

Postcard mailed to lawyers and/or insurance companies named by mail or telephone respondents

tion brought in either a Circuit or Federal District Court. Second, if the completed postcard stated that the individual inquired about was a plaintiff in a Circuit or Federal District Court case, this was considered conclusive and the individuals represented by these cards became the nucleus of the police sample group that would be dropped in favor of the court sample. Third, if the name of a lawyer or insurance company was not given by the original respondent, or if the card was not returned by the lawyer or insurance company, or if the lawyer or insurance company could not provide the information requested, a form letter and postcard (similar to the card sent to lawyers) was sent to the original respondent. The cover letter thanked the individual for having previously completed either the mail or telephone questionnaire and explained that the additional information was needed in order to develop a complete statistical picture of the results of automobile accidents in Michigan. The postcards returned by this group were treated in the same manner as those returned by the lawyers or insurance companies.

The results of the mailing (at a cut-off date two weeks after the last letter had been mailed) are shown in Table 9-8. The fact that the number of plaintiffs exceeds the number of defendants might be explained in a number of ways. First, it could be argued that a defendant who is also the plaintiff in a counterclaim would be more likely to return the postcard marked "plaintiff." Second, for accidents involving a number of individuals or vehicles, there are often more plaintiffs than defendants, i.e., if an individual is generally considered to be at fault, it is probable that he will be sued by a number of other persons. Third, it might be expected that mail or personal inquiries would present less of a psychological threat to plaintiffs than to individuals who have been accused of being responsible for the accident. And finally, although there may be a fifty-fifty chance that an injured person was at fault, there is a considerable chance that he did no damage to the other party, either because collision was with a fixed object (e.g., tree, ditch, or parked car) or because the other vehicle was less vulnerable (e.g., truck or railroad). Also it is well known that a seriously injured person is less likely to be sued because (1) the jury will sympathize with him, or (2) he is impoverished by the accident and can't pay.

TABLE 9-8

Designation of Plaintiffs in Police Sample: Results of Mailing to Lawyers, Insurance Companies, and Initial Respondents

Results of mailing	*Number of cases*	*Percent of sample*
Postcard questionnaire was returned	*370*	*51%*
Plaintiff in Circuit or Federal Court	38	5
Defendant in Circuit or Federal District Court	20	3
Party to a suit brought in a lower court	21	3
Not a party to litigation	256	35
Lawyer or insurance company could not provide the information requested	34	5
Postcard questionnaire was not returned	*355*	49
Total	724	100%

Table 9-8 shows that in 34 cases the postcards were returned incomplete, and in an additional 355 cases, the postcards were not returned. To effect the desired substitution, all 389 individuals asked about on these questionnaires had to be "assigned" to one of the three following categories: (1) plaintiff in Circuit or Federal District Court, (2) defendant in Circuit or Federal District Court, or (3) party to a suit brought in a lower court or not a party to any litigation.

The assignments were made by first examining all available information on each case, and where it seemed highly likely that the individual had not been a party to litigation, the individual was assigned to the third category.[5] (Ten individuals were assigned using this procedure.) Next, on the basis of information available from the entire study, criteria were selected that could be used to associate an individual with one of the three categories. For example, all individuals arrested or cited for a violation were assigned to the "defendant" category on the basis of the fact that the large majority of defendants who did return a postcard questionnaire had been arrested or cited for a traffic violation. Insofar as possible, a number of variables were used to assign each case, and variables that were best able to discriminate among the three categories were given more weight in classifying the cases. This assignment procedure was continued until every individual had been included in one of the three categories. Of the 2872 individuals in the police sample 105, or 3.7 percent of the sample, were designated as plaintiffs in a Michigan or Federal District Court. Thirty-eight of these had returned the postcard questionnaire. Eleven were designated plaintiffs as a result of information available from the personal interviews, and the remaining 56 were assigned using the procedure just outlined.

As indicated on Figure 9-1, for analysis purposes the 105 plaintiffs from the police sample were dropped, and the plaintiffs from the court sample were substituted for them. Both the logic underlying this type of sample substitution procedure and the effectiveness of the substitution in this particular study are discussed in Chapter 10.

---

[5] Supplement A to this report describes the assignment procedure in more detail than is necessary here.

I. PERSONAL INTERVIEWS WITH CLAIMANTS' LAWYERS

1. *Purpose*

The claimants' lawyers' study was the first study completed under a grant from the Walter E. Meyer Research Institute of Law. As specified in the research proposal, the additional grant was obtained to pursue three principal objectives—first, to obtain data on heretofore unstudied questions, such as the legal problems that hinder prompt settlement of claims; second, to verify financial information already obtained from the initial personal interviews; and third, to provide information on methodology which would enable others to make future surveys of auto accident compensation with increased efficiency at a lower cost. The claimants' lawyers' study, along with the studies that will be described in the immediately following sections of this chapter, was designed to be combined with data obtained in the first part of the study to achieve these objectives.

2. *Questionnaire Content and Eligible Respondents*

Claimants' lawyers were questioned about legal actions taken on behalf of their clients as well as about the costs and compensation to both themselves and their clients. Following a detailed discussion of the specific case sampled, they were asked a number of general questions concerning their views about the way auto-injury cases presently are handled in Michigan and about the problem of court delay in Michigan.

The claimants' lawyers' sample was designed so that it could be treated as an independent study or combined with other parts of the total study for case analysis. To be more specific, claimants' lawyers were selected from two sources. First, all the plaintiffs' lawyers listed on the sampled court calendars were included in the sample, regardless of whether or not an interview had been completed with their client. Second, lawyers hired by personal-interview respondents from the police sample were included only if their clients had given permission, either verbally or in writing, for the interview to take place. The composition of the claimants' lawyers' sample is shown in the first column of Table 9-9.

A further word should be said about the request for permission to interview lawyers. Each of the 378 personal-interview respondents who hired a lawyer was asked to sign a waiver of confidential privilege giving permission for his lawyer to disclose costs and legal issues in the case.[6] If the personal-interview respondent was reluctant to sign the waiver, he was asked to give verbal permission for his lawyer to be interviewed. And if either verbal or written permission was given, he was asked for the name and address of his lawyer. Seventy-one percent of those asked to sign the waiver did so, and 87 percent of the lawyers listed on the waiver agreed to be interviewed. This compares with an overall response rate of 77 percent for all lawyers. (The precise sequence of questions is shown in Appendix B, questions F-14 to F-16.) A detailed evaluation of the effectiveness of waivers in securing interviews with lawyers is presented later in this chapter.

Returning to Table 9-9, consider again the two groups of eligible respondents. The 207 plaintiffs' lawyers who were listed on the sampled court calendars comprise a probability sample representative of all lawyers filing on behalf of plaintiffs in Michigan personal-injury automobile accident suits during one calendar year. For purposes of estimating state aggregates, each of these cases need only be weighted in accordance with both its initial

---

[6] The use of a waiver was suggested by members of the advisory committee at a meeting of the committee in April, 1960. The committee felt that a signed release from a lawyer's client would substantially increase the overall response rate for this part of the study. The waiver was pretested in interviews with injured individuals or members of their families, and its use seemed to offer no significant interviewing problems.

Alfred F. Conard

TABLE 9-9

Sample Sizes and Response Rates for Personal Interviews Completed with Claimants' Lawyers

Sample source	*Number of lawyers sampled*	*Number of questionnaires completed*	*Number of questionnaires not completed*	*Percent of questionnaires completed*
Plaintiffs' lawyers listed on sampled court calendars	207	149	58	72.0%
Lawyers hired by personal-interview respondents from the police sample who granted permission for the lawyer to be interviewed	63	58	5	92.1
Total sample	270	207	63	76.7%

chance for inclusion in the sample (i.e., the sampling interval used to select the court calendars) and the personal-interview response rates. The response rates and determination of interview weights for this group are shown in Table 9-10.

The second group of eligible respondents shown in Table 9-9, the 63 lawyers hired by personal-interview respondents from the police sample, does not comprise a representative sample of any larger group. Therefore, the 58 completed questionnaires from this group have not been weighted nor have they been included in any of the quantitative material presented in earlier chapters; they have been used only for individual case studies. The quantitative material in this report is based entirely on the 149 questionnaires completed with the plaintiffs' lawyers listed on the sampled court calendars.

3. *Results*

The response rates for the two claimants' lawyers' groups are shown in Table 9-9. The questionnaire completion rate is noticeably higher for the lawyers hired by personal-interview respondents from the police sample than for plaintiffs' lawyers listed on the sampled court calendars. A comparison of the two nonresponse groups explains the difference. Table 9-11 shows why interviewers were unable to complete interviews with 58 of the plaintiffs' lawyers listed on the court calendars. Notice that twenty-six of the nonresponse interviews (44.8 percent) were never attempted. Nineteen of these "automatic nonresponse" were lawyers whose clients had either refused to be interviewed or had asked that no interview be attempted with their lawyers; although it was felt that interviews should not be attempted with the lawyers in these cases, the cases were included in the statewide plaintiffs' lawyers' sample by definition.

Ten additional plaintiffs' lawyers declined to complete the questionnaire because they could not remember details of the case and the case file could not be located. For all 10 of these cases,

Alfred F. Conard

TABLE 9-10

Response Rates and Determination of Interview Weights for Information from Court Calendars and for Personal Interviews Completed with Plaintiffs' Lawyers from the Court Sample

Strata									
Time required to settle case	*Size of suit*	*Initial sampling rate for court cases (1)*	*Number of court cases sampled (2)*	*Weight for case information from the court calendar (3)*	*Sub-sampling rate (4)*	*Number of court cases in subsample (5)*	*Number of plaintiffs' lawyers selected for personal interview (6)*	*Percent of personal interviews completed (7)*	*Weight of each completed personal interview (8)*
				Circuit or Kent Superior Court					
2 years or more	More than $25,000	1/16	55	16	1	55	55	72.7%	22
	$25,000 or less	1/16	63	16	½	31	31	80.6	40
Less than 2 years	More than $25,000	1/16	43	16	1	43	44[a]	68.2	23
	$25,000 or less	1/16	84	16	½	40	40	62.5	51
				Federal District Court					
2 years or more	More than $25,000	¼	13	4	1	13	12[b]	91.7	4
	$25,000 or less	¼	11	4	½	5	5	100.0	8
Less than 2 years	More than $25,000	¼	12	4	1	12	12	66.7	6
	$25,000 or less	¼	14	4	½	8	8	62.5	13
Case still open			14			0	0	0.0	
Total sample			309			207	207	72.0%	

[a] Court case No. 8112 was treated as two separate interviews.

[b] Court case No. 8131 was removed from sample (case reopened).

TABLE 9-11

Reason Why Questionnaire Was Not Completed with Plaintiff's Lawyer, Court Sample

Reason why questionnaire was not completed	*Number of questionnaires not completed*		*Percent of questionnaires not completed*
Interview never attempted (automatic nonresponse)		26	44.8%
Plaintiff refused to be interviewed	11		
Plaintiff asked us not to talk with his lawyer	8		
Respondent (plaintiff's lawyer) is ill or deceased, or moved out of state, or current address could not be determined	5		
Other reasons	2		
Interview refused by respondent (by reason given):		18	31.0
Respondent too busy; cannot afford time	4		
Respondent will not complete interview until a release from his client is obtained	8		
Respondent presently under suspension by Michigan Bar	2		
Case still open	1		
Other reasons	3		
Interview prevented by unavailability of respondent's case file (by time between accident and interview):		10	17.3
Less than four years	—		
Four years, but less than five years	2		
Five years, but less than six years	6		
Six years or more	2		
Respondent absent (out of state or country) during interviewing period (no one else available to complete questionnaire)		4	6.9
Total questionnaires not completed		58	100.0%

the accident inquired about had occurred at least 4 years prior to the interview. The two groups just discussed account for 62.1 percent of the nonresponse. The remaining 37.9 percent of the nonresponses are classified by reasons common to all field surveys —they were either refusals or cases where the designated respondent was absent from the state or country during the interviewing period.

For comparison, a number of explanations might be suggested for the extraordinarily high proportion of questionnaires that were completed with lawyers hired by personal-interview respondents from the police sample. First, compared with the court sample, the recentness of the accidents considerably reduced the chance of the file not being available or of the lawyer not being able to remember the case. Second, in all the cases included in this group, the lawyer's client had previously been interviewed and the client had given his verbal or written consent for his lawyer to be interviewed. It will be shown shortly that securing permission from the client did increase the response rate with lawyers. A summary of the reasons given for the five nonresponses in this latter group is shown in Table 9-12.

TABLE 9-12

Reason Why Questionnaire Was Not Completed with Claimant's Lawyer, Police Sample

Reason why questionnaire was not completed	*Number of questionnaires not completed*
Interview never attempted (automatic nonresponse)	2
Interview refused by respondent	1
Interview prevented by unavailability of respondent's case file	1
Other (respondent is not an attorney, although he did give claimant legal advice)	1
Total questionnaires not completed	5

4. *A Note Concerning the Effectiveness of Waivers*

Tables 9-13 and 9-14 indicate that the securing of a client's permission to interview his lawyer does substantially increase the probability of completing an interview with his lawyer. In Table 9-13 the bottom row (which shows the response rate for interviews that were assigned to the field, i.e., the results for all cases where an interview was actually attempted) indicates that about 90 percent of the interviews were completed when the interviewer had a signed waiver available to show the lawyer, that about 87 percent of the interviews were completed when the interviewer could state that the client had given his verbal permission for the interview to take place, but that only about 72 percent of the assigned interviews were completed when the client's permission had not been secured.[7]

Looking at both Tables 9-13 and 9-14, it appears that the fact of having permission is more important than the form in which the permission is given. There is essentially no difference between the response rates for lawyers who were presented with a signed statement and lawyers who were simply told that their client had granted permission for the interview. Moreover, Survey Research Center interviewers were instructed to leave the waiver with the lawyer if he so requested; most lawyers gave the statement merely superficial attention and returned it to the interviewer. From the evidence presented here, one would conclude that securing a claimant's permission to interview his lawyer did contribute to the overall response rate; but the evidence does not suggest that the use of a waiver produced significantly better results than did the securing of verbal permission only.

---

[7] It should be kept in mind, however, that 8 of the 31 nonrespondents in this third group (See Table 9-11.) told the interviewer that they would complete a questionnaire if the signed release could be obtained. It is entirely possible that these 8 attorneys might have completed the questionnaire if the issue of waivers had not been emphasized in both the introductory letter to the lawyer and in the introduction to the personal interview. If this had been the case, the response rates in the three groups would have been almost equal. On the other hand, it is equally possible that lack of permission in one form or the other could have substantially reduced the response rate in the first two groups.

7.

Alfred F. Conard

TABLE 9-13

Personal Interview Results with Plaintiff's Lawyer by Personal Interview Results with Plaintiff and by Whether Plaintiff Signed a Release, Court Sample

		Personal interview results with plaintiff			
		Questionnaire completed			
		Permission to interview lawyer granted			
Personal interview results with plaintiff's lawyer	*All cases*	*Release signed*	*Release not signed*	*Permission to interview lawyer refused (interview never assigned)*	*Questionnaire not completed*
Number of lawyers' questionnaires	207	86	32	8	81
Number of lawyers' questionnaires completed	149	73	26	—	50
Number of lawyers' questionnaires not completed	58	13	6	8	31
Not completed because interview never assigned; (See Table 9-11)	26	5	2	8	11
Percent of questionnaires completed (all cases)	72.0%	84.9%	81.3%	0.0%	61.7%
Percent of questionnaires completed (excluding interviews never assigned)	82.3%	90.1%	86.7%	—	71.4%

TABLE 9-14

Personal Interview Results with Claimant's Lawyer by Whether Claimant Signed a Release, Police Sample

Personal interview results with claimant's lawyer	*All cases*	*Whether claimant signed a release*[a] *Release signed*	 *Release not signed*
Number of lawyers' questionnaires sampled	63	48	15
Number of lawyers' questionnaires completed	58	44	14
Number of lawyers' questionnaires not completed (for any reason)	5	4	1
Percent of questionnaires completed (all cases)	92.0%	91.7%	93.3%

[a] All respondents granted permission for Survey Research Center interviewers to talk with their lawyers about the case.

J. TELEPHONE INTERVIEWS WITH DEFENDANTS

1. *Introduction*

A second empirical study completed under the Meyer Grant involved the completion of telephone questionnaires with defendants listed on the sampled court calendars. In developing the research design for this part of the project, a number of problems were encountered, the most important of which are discussed briefly here.

Corporations named as defendants were excluded from the sample for a number of reasons. First, the interviewing schedule designed to be completed with individual defendants was not appropriate for interviewing corporation officials, and there were not enough corporations in the sample to justify a separate analysis. A less important reason involved the practical difficulty of locating the individual within a corporation's claims office who handled a particular case. More important, if the individual could be located, the information asked of him would have been essentially the same as that requested on the defendants' lawyers' questionnaire for the same case. In fact, in many cases the corporation's claims representative and the defendant's lawyer would have been the same individual. Governmental units named as defendants were also excluded from the sample for similar reasons.

Multiple defendants offered an additional design problem. The unit of analysis had been defined as all defendants listed on the court calendars. For cases where two or more related individuals living together were named as joint defendants, such as a married couple, the interviewer was instructed to complete first a regular questionnaire with either of the individuals named as a defendant. A shorter supplemental questionnaire was then completed for each additional family member named as a defendant. The supplemental form repeated only those questions from the regular questionnaire for which the answers would be expected to vary between members of the same family, such as whether the person was driving at the time of the accident and the basic demographic data. Interviewers were permitted to complete the supplemental questionnaires with either the original respondent or with the family member being asked about. A review of the interviewing results shows that all the supplemental questionnaires were completed with the original respondents.

If a defendant had died between the time the suit was filed and the interviewing period (regardless of the cause of death), the interview was assigned as an automatic nonresponse; that is, no alternative respondent was allowed. In suits where the defendant was the administrator of an estate, the interviewers were instructed to alter the wording of selected questions on the regular questionnaire by hand, so that the questions asked administrators would be consistent and appropriate.

Use of the most economical data-collection technique was also given careful consideration. A pretest in Detroit indicated that the schedule could be completed by telephone with little difficulty, and it was decided to attempt the initial defendants' interviews by telephone, with the option of switching to personal interviews if any serious difficulties were encountered in the telephone procedure. Early interviews were assigned with out-of-state respondents who, in most cases, would have been assigned as automatic nonresponses using personal-interview techniques. These interviews were successful, and a second group of interviews were assigned with respondents in outlying areas of Michigan (these respondents would have represented high-cost personal interviews), and finally the entire sample was assigned for telephoning.

2. *Questionnaire Content and Eligible Respondents*

The defendants' questionnaire was designed to examine the process by which respondents defended themselves. Every effort was made to separate aspects of the defense process from other aspects of the accident, such as medical expenses, property damage, etc., incurred by the defendant or his family. An effort also was made to focus the respondent's attention on issues relevant to the particular suit sampled, as opposed to other suits that might have been brought against the defendant or his family as a result of the same accident. The questionnaire itself asked about the defendant's involvement in legal proceedings (including the manner in which he secured counsel), and about any psychological and economic penalties incurred by himself or his family as a result of the suit.

3. *Results*

The response rates and the determination of interview weights for completed defendants' questionnaires are shown in Table 9-15. Note that even with the exclusion of private corporations and governmental units, there were 267 individual defendants listed on the 207 sampled court calendars. The same 207 calendars listed 207 plaintiffs (Table 9-7); thus, the court calendars listed about 1.29 defendants for each plaintiff.

The mobility of American families, particularly younger families, increases the problem of tracing respondents when the respondents (and their addresses) are selected from a list that is not current. Table 9-16 indicates clearly that the effectiveness of the procedures used to locate defendants was important in establishing the over-all response rate for this part of the study. Of the 267 defendants designated for personal interviews, 55 (65.5 percent of the nonresponse group) could not be located; no interviews could be attempted with this group.[8] Table 9-16 also shows that only 8 of the defendants refused to be interviewed by telephone.

---

[8] The reader may be interested in a brief description of the procedures used to locate sampled defendants. An initial effort was made to find some known address for each defendant, regardless of how old it was. Court records provided complete 1957 addresses for about one-half the sample; however, for most of the remaining defendants, the 1957 city of residence was mentioned somewhere in the court records. Current (1961) telephone books were checked next, and if the 1957 address found in the court records and the address shown in a current telephone book were the same, the questionnaire was assigned for interviewing. For cases where the court records indicated only a city of residence, the current telephone book for that city was checked, and if an individual with the same name as the defendant was listed in the telephone book, and there was some reason to be

K. MAIL INTERVIEWS WITH DEFENDANTS' LAWYERS

1. *Introduction*

A final empirical study required interviewing all defendants' lawyers listed on the sampled court calendars, using a two-page mail questionnaire. This study was undertaken so that the views of all major participants in the litigation process would be included in the final report, thus permitting comparisons on a case-study basis. Although the personal-interview questionnaire used with plaintiffs' lawyers is much more detailed than the mail questionnaire used with defendants' lawyers, the nine questions included on the defendants' lawyers' questionnaire were also asked of the plaintiffs' lawyers. The questions are primarily concerned with the major issues or sources of disagreement in the case and the important factors underlying determination of a final settlement.

The principal factor which limited the length of the defendants' lawyers' questionnaire and dictated the use of a mail questionnaire was the concentration of a large percentage of the defense work in a relatively few firms. In most instances, these firms are affiliated with one of the insurance companies writing large amounts of automobile insurance in Michigan. Table 9-17 compares the concentration of defense lawyers with the concentration of plaintiffs' lawyers for the 207 sampled court cases. It is quickly apparent that a relatively few law firms handle a large propor-

---

confident that the telephone listing was that of the defendant (e.g., if there was only one individual with the defendant's name listed in the telephone book, or if a telephone book listing showed the same last name and the same address as shown on the court records, but perhaps a different first name), the questionnaire was assigned for interviewing.

For defendants still not located, additional efforts were made to find the police report for the accident (which might list either the the address of the defendant or the name and address of an individual who would know the present whereabouts of the defendant). Further investigation was carried out at the R. L. Polk Library in Detroit and at the Michigan Drivers' License Bureau in Lansing. Some success was achieved through each of the sources mentioned.

TABLE 9-15

Response Rates and Determination of Interview Weights for Information from Court Calendars and for Telephone Interviews Completed with Defendants from the Court Sample

Strata									
Time required to settle case	*Size of suit*	*Initial sampling rate for court cases (1)*	*Number of court cases sampled (2)*	*Weight for information from the court calendar (3)*	*Sub sampling rate (4)*	*Number of court cases in subsample (5)*	*Number of defendants selected for personal interview[a] (6)*	*Percent of personal interviews completed (7)*	*Weight of each completed personal interview (8)*
				Circuit or Kent Superior Court					
2 years or more	More than $25,000	1/16	55	16	1	55	77	68.8%	23
	$25,000 or less	1/16	63	16	½	31	42	76.2	42
Less than 2 years	More than $25,000	1/16	43	16	1	43	64	68.8	23
	$25,000 or less	1/16	84	16	½	40	51	70.6	45
				Federal District Court					
2 years or more	More than $25,000	¼	13	4	1	13	13	38.5	10
	$25,000 or less	¼	11	4	½	5	5	100.0	8
Less than 2 years	More than $25,000	¼	12	4	1	12	11	72.7	6
	$25,000 or less	¼	14	4	½	8	4	0.0	—
Case still open			14			0	0	0.0	
Total sample			309			207	267	68.5%	

---

[a] Private companies and governmental units were excluded from this sample. (There are 15 such cases in the 38 cases sampled from the Federal District Courts.)

TABLE 9-16

Reason Why Questionnaire Was Not Completed with Defendant, Court Sample

Reason why questionnaire was not completed	*Number of questionnaires not completed*	*Percent of questionnaires not completed*
Respondent refused to be interviewed (does not want to talk about it; too upset; too busy; can't remember details)	8	9.5%
Respondent ill or deceased (no one else available to complete questionnaire)	16	19.0
Respondent absent (out of state or country during interviewing period; no one else available to complete questionnaire)	5	6.0
Respondent could not be located (unable to determine current address or telephone number)	55	65.5
Total questionnaires not completed	84	100.0%

tion of all defense work for Michigan automobile suits. To be more precise, the survey indicates that five law firms handle 38 percent of all the defense work resulting from Michigan personal-injury automobile suits. (One law firm accounted for 36 percent of the total nonresponse for this part of the study.) With this heavy concentration of defense work, it seemed unreasonable to expect that the lawyers involved would be willing, or should be asked, to complete more than a few questions about any one case. These considerations led inevitably to the conclusion that the final questionnaire should be short and that the respondent should be allowed to complete the schedules as his time permitted. Of course, if it had not been essential to interview about the specific

TABLE 9-17

Concentration of Cases for Sampled Plaintiffs' and Defendants' Lawyers, Court Sample

Number of sampled cases handled by lawyer or law firm	*Whether lawyer or law firm represented plaintiff or defendant*	
	Plaintiffs	*Defendants*
1	87%	72%
2—3	11	15
4—6	2	7
7—9	0	1
10—12	0	2
13—15	0	1
16—24	0	1
25 or more	0	1
Total	100%	100%
Total number of lawyers or law firms in sample	215[a]	102
Range of number of cases handled by single lawyer or law firm	1—5 cases	1—28 cases

[a] A small number of these lawyers were not listed on the 207 sampled court calendars, but were interviewed when it was subsequently determined that they played a major role in settling the case.

cases in the court sample, the sample would have been designed to avoid the extreme clustering of interviews evident here. But an efficient representative sample would still be concentrated to some extent in certain law firms because the cases are concentrated in these firms.

2. *Results*

The response rates and determination of interview weights for interviews completed by defendants' lawyers are shown in Table 9-18.

57.p

L. ESTIMATION OF ECONOMIC LOSSES

1. *Introduction*

For the purposes of this study, comparing losses with compensation, it is the loss to the individual or his family which is relevant,

A SAMPLE WITHIN A SAMPLE

Previous discussions of survey techniques of gathering data have involved examples where the universe under observation has been conveniently contained in some sort of official record or list. Thus Zeisel referred to Census data which is collected by interviewing "every fifth household," and Conard's study elicited information from a universe defined by "police accident reports."

In the following reading, however, research employing the survey technique is discussed in relation to a universe which cannot be defined by readily accessible lists of people, things, or events. The Gallup Poll is concerned with predicting election results. The universe, consequently, is defined as the voting public of the United States. And this means that the Poll is concerned not just with the adult population in a particular community or state but with those who in fact will vote in the country as a whole. And, of course, there is no list containing the names of all such people in the United States.

The solution to this problem is to engage in a multi-stage sampling process. The universe is divided into small manageable units which may be sub-divided in turn. As Paul Perry indicates: "A systematic sample of cities and minor civil divisions is drawn . . . . For election survey purposes a selection of smaller units . . . is drawn . . . . In cities, such smaller units are usually wards. Within these units precincts are selected."

It must be remembered that the success of this approach, is dependent upon preserving the basic idea of random sampling, namely, that every unit finally chosen has an equal probability of being selected. It is only when this assumption is valid that we are able to use much of the statistical theory basic to statistical inference--an issue we will explore in the next chapter.

From Paul Perry, "Election Survey Procedures of the Gallup Poll," The Public Opinion Quarterly, 24 (Fall, 1960), pp. 531-542. Reprinted by permission.

Offprint from THE PUBLIC OPINION QUARTERLY
Volume 24, Fall 1960

Paul Perry, "Election Survey Procedures of the Gallup Poll," The Public Opinion Quarterly, Vol. 24, Fall, 1960, pp. 531-542.

The development of sample surveys of the population has been accompanied by their highly publicized use as a means of determining the voting public's political preferences. The use of sample surveys for this purpose has provided a difficult test of the method, particularly in the United States, where a large part of the adult population fails to vote. Testimony to this effect appeared in Chapter I of *The Pre-election Polls of 1948,* a Social Science Research Council publication issued in 1949. The statement was: "The committee and the staff would like to point out that pre-election polling and predictions constitute one of the more severe tests of polling or survey methodology. There are more steps in pre-election polling at which error can enter than there are in many other applications of survey technique."

Since a straightforward sample survey of the total adult population on opinion with respect to parties or candidates quantifies only one variable in the formula, this has provided the Gallup Poll with some difficult problems. What is specifically required, of course, is an estimation procedure that takes into account the major variables and conditions. Progress in the development of such an estimation procedure by the Gallup Poll has increased the reliability of its pre-election estimates of the division of the vote based on survey data. For this reason it probably has some intrinsic interest for many who are engaged in survey research, and the methods used may prove useful in other political research.

The methods of sampling the voting population used, relating voting participation to candidate preferences, dealing with the problem of the undecided, obtaining unbiased expressions of candidate preference, and measuring reactions to major events close to election day are of methodological concern to all engaged in political research.

In this discussion data reported concerning election surveys and samples are based on the United States excluding the South.[1] The

---

* The author is president of the Gallup Organization, Inc., and research director of the Gallup Poll.

[1] States excluded: Alabama, Arkansas, Florida, Georgia, Kentucky, Louisiana, Mississippi, North Carolina, Oklahoma, South Carolina, Tennessee, Texas, and Virginia.

South has been excluded because it presents special difficulties which introduce additional variables. For example, in Congressional elections many districts in the South are uncontested. Another difficulty is presented by the fact that the South's share of the total vote fluctuates between elections to a much greater extent than is true for other major regions. The limited scope of the article precludes going into the details of the methods of coping with these factors.

The data reported are based on Gallup Poll experience in the elections of 1950, 1952, 1954, 1956, and 1958. In 1950 certain of the basic elements in the estimation procedure here described were first adopted and used. We believe it is noteworthy that as a result of using these methods, our pre-election estimates of how the popular vote divided in four of these elections (1950, 1954, 1956, and 1958) were the four most accurate estimates of the twelve which the Gallup Poll has made in national elections. The first of the twelve estimates was made in the 1936 presidential election.

THE SAMPLE

The sample of areas used for the Gallup Poll election surveys is the same as the sample of areas used for the Gallup Poll surveys between election periods. These areas are drawn in accord with the principles of the theory of probability sampling. Between election periods the areas consist of block clusters and rural segments. In the two election surveys which provide the basis for the final pre-election estimate, the areas consist of election precincts, drawn in the same manner as the block clusters and rural segments. The change to election precincts in the period immediately preceding the election is made because of the greater control provided with regard to the political representativeness of the sample of areas. For surveys on most general issues the precinct as the basic areal sampling unit would have little or no advantage over block clusters and segments and might have some disadvantages.

The sample of areas is drawn in the following manner: A systematic sample of cities and minor civil divisions is drawn from regional-city size strata with probability of selection proportional to size. Within places so drawn, for election survey purposes a selection of smaller units for which election data are available is drawn in the same manner. In cities such smaller units are usually wards. Within these units precincts are selected. The selection of the precincts proceeds in this manner: Election results for the previous national election are obtained for each precinct in the ward. One precinct is then drawn with probability of selection proportional to the precinct's total vote. Within the precinct a systematic sample of households is selected, and one adult from each household is interviewed. The total number of adults in each household is obtained in the survey to provide for a size-of-household correction when necessary.

For each precinct so selected, the proportions in which the vote divided by parties in the previous one or two national elections is computed. In 1958, for example, the percentage voting for Eisenhower of the Eisenhower-Stevenson vote in 1956 was computed for each precinct drawn into the sample. Since the precincts were drawn with probability proportional to size, the mean Eisenhower percentage of the precincts in the sample served as a measure of the representativeness of the areas drawn. For example, if the mean Eisenhower percentage for such a sample was 58.5, the sample of precincts was 0.7 percentage points higher than the vote for Eisenhower in 1956, when, of the vote for Eisenhower and Stevenson, 57.8 per cent voted for Eisenhower. When such a bias is found to exist, a simple correction of the final estimate can be made, and the bias in the sample of precincts with respect to the previous election removed.

The vote in a previous election is perhaps the most efficient control one can use for an election survey (unless the span of time between elections is very great), either as a basis for stratification or as a part of the estimation procedure. However, difficulties are involved in using it as a basis for stratification in a national election, at least if precincts are the areal units. Precinct data are not available from a central source; a great deal of time and effort must be expended to obtain election results for even a modest number of precincts. The use of precinct data becomes quite laborious even with a double sampling procedure, in which a large sample of precincts is drawn first, strata are formed, and the final sample is drawn from the strata. Therefore, it appears to be more efficient to use the past voting data in forming the estimates after the sample is taken instead of using them in drawing the sample.

The correlation of the vote by precincts between elections is fairly high. For example, we have found the variance between precincts in terms of shift or change in the percentage of the vote for a candidate of a given party between two elections is only about two-thirds as large, typically, as the variance between them in terms of party or candidate preference in a single election.

In the 1950, 1952, and 1956 samples the deviation of the mean precinct party percentage for the previous election used as a control was small; in 1954 and 1958 it was of some consequence. The deviations of the means for each previous election are shown in Table 1.

This step cannot be expected, of course, to improve matters in every instance. Some of the time it will make things worse. In view of the

Paul Perry

TABLE 1

DEVIATIONS OF THE MEANS OF THE SAMPLES OF PRECINCTS* IN PER CENT REPUBLICAN OF THE DEMOCRATIC-REPUBLICAN VOTE

Survey Year	*Past Election Used*	*Deviations*†	*Number of Precincts*
1950	1946	+0.7	50
1952	1948	+0.3	100
1954	1952	+2.8	105
1956	1952	−0.2	171
1958	1956	−2.2	156

* Outside the South.

† For example, the 1946 election results in precincts in the sample in 1950 had a mean per cent Republican vote 0.7 of a percentage point higher than was true in the United States outside the South. As a sample of precincts, they were this much in error in terms of the results of the previous election.

correlation between elections in the division of the vote, however, it can be expected to improve matters at least the majority of times.

The above illustrates one of the advantages of using precincts as the sampling unit as compared with other areal units such as census tracts or enumeration districts. While it is not possible to obtain for precincts census statistics such as those concerning demographic characteristics of the population, this disadvantage is more than offset in our opinion by the availability of election data. In fact, the availability of census data by small areal units is of limited value in any case, because it is useful only in the first few years following the decennial census.[2]

Within the precinct, unless it has an exceedingly large population, the sampling interval used in the selection of households requires the interviewer to cover the area entirely. That is to say, the sample does not, in most cases, consist of clusters of households. The interviewer is given a randomly selected starting point and takes every *n*th household throughout the precinct. In larger precincts, a sample of blocks or segments is selected and, within the blocks or segments drawn into the sample, a systematic sample of households is drawn as above. These are precincts where the sampling interval for coverage of all blocks in the precinct would be so large that it would create serious problems of maintaining an accurate count.

Within the household the individual who is interviewed is selected from among those who are at home, in the following manner: The interviewer asks first to speak to the youngest man of voting age who is at home; if no man is at home he asks to speak to the youngest woman who is at home. If no one is at home or the indicated person refuses, the interviewer is told to go to the next adjacent dwelling unit. In such a case the interviewer resumes the interval count from the original dwelling selected. The interviewer is given a male-female assignment. Thus if, following the procedure described above, the number of men assigned is obtained, in subsequent households the interviewer asks to speak only to the youngest woman at home. Interviewing hours are from 4 P.M. to 10 P.M. A record of refusals is kept, with information concerning the person's sex, estimated age, education, race, and whether the resident is in a high-, medium-, or low-rental area.

---

[2] The use of precincts as the areal sampling unit in election surveys also has some practical advantages over use of other divisions such as census tracts or enumeration districts. For example, if interviewer failure causes loss of a sampling area too late for replacement, knowledge of the previous voting behavior of the area enables one to adjust statistically for any resultant political bias.

A point commonly made about the use of precincts as sampling units is that, since the boundaries of many of them are frequently changed, their usefulness is limited as a political control. This objection is based on a mistaken notion of how they are used. The vote in a previous election is the control, and therefore the boundaries of the precincts as they were at the time of that election are the ones used. It makes no difference whatsoever how frequently they are changed thereafter. It would be useful, of course, in analyzing the accuracy of the survey results if the boundaries were constant from election to election. We have found that, on the average, about 75 per cent of precincts remain the same over a four-year period, and we use these for error analysis.

LATE CAMPAIGN SURVEY

A discussion of the use of the precinct as the ultimate areal sampling unit in the design leads into a description of the method used to bring the measurement of voting intentions as close to the election as possible. The use of precincts on a large scale by the Gallup Poll began with the 1950 congressional election. Although a method of carrying out a late October survey and getting the results in time for pre-election publication had been worked out and used in 1940 and 1944 (unfortunately *not* in 1948), a less cumbersome method was sought. A combination of the telegraphic communication of survey results to the home office with the control provided by the use of precincts seemed to be particularly advantageous and was the method adopted. The two final election surveys are usually conducted in October. The first one usually takes place the first week in October; the second on Wednesday, Thursday, and Friday of the week prior to the election. The first survey questionnaires are returned by mail. However, the interviewers are also required to carry out all the steps involved in the final telegraphic survey, including tabulation of the results, recording and checking the tabulations on forms provided, and preparation of the telegram. We find that about 80 per cent of the interviewers do this quite precisely and accurately, about 15 per cent make small errors which cannot seriously affect the results, and the remainder make serious errors which have to be dealt with. In the latter case, the questionnaires are returned to the interviewers and they are requested to read their directions once again, do over the tabulation, and record the corrected data. Telephone consultations are then conducted with each of them to be sure their confusion has been eliminated. Those who make minor errors are also told about it, and once again great stress is laid on the necessity of accuracy. A simplified breakdown of the data has to be employed, of course, to minimize the possibility of errors. For example, interviewers are required only to divide their questionnaires into two easily classified groups—one group consisting of those who say they plan to vote, the second group comprising all others. They tabulate the voting preference only among those who say they plan to vote. Interviewers are also asked to tabulate and telegraph the educational attainment of all respondents in six categories, but not the *vote* by education categories. The latter step would increase the chance of transmission errors so much that it is not considered worth the risk.

As a variable closely associated with socio-economic status and income, the educational-attainment information is used to standardize the two surveys in this respect. Since the vote by education is not obtained, for the reason indicated above, any difference between the two samples by education is dealt with in this manner: The first survey is tabulated by the six education categories (college complete, college incomplete, high school complete, high school incomplete, grammar school seventh and eighth grades, grammar school sixth grade or less), and the vote or preference in each category. The data are then adjusted to fit the education distribution of the second sample, and the change in voting preference which results is applied to the results of the second survey. This sample control of the second survey is designed primarily to protect against the consequences of interviewer failure to do a careful coverage of the precinct the second time around. Since this control has been employed, the effect of the adjustment has been in no case more than one- or two-tenths of a percentage point.

TURNOUT

Another important step in the procedure for estimating election results is projecting the proportion of the population old enough to vote who will vote, or what might be called the turnout ratio. Our estimate of this ratio is derived from the results of several questions which are related to voting participation. Some combinations used, for example, are based on these two questions: "How much thought have you given to the coming November elections—quite a lot, or only a little?" and, "Do you, yourself, plan to vote in the election this November, or not?" If the answer is "yes," the respondent is asked, "How certain are you that you will vote—absolutely certain, fairly certain, or not certain?" For previous elections a record has been kept of the relationship between the proportion who answer such questions in a category indicating a likelihood of voting and the actual turnout ratio in the election. From these relationships a sufficiently accurate estimate of the turnout ratio can be made for estimating preference in that part of the sample representative of those who will vote. For example, the method projected a turnout percentage of 61 in 1956 as compared with the actual turnout ratio of 60; in 1958 it projected a turnout of 45 as compared with the actual turnout of 44. In 1950, 1952, and 1954, the turnout ratio was projected from available registration data.

The variable of voting participation is an important one in United States elections, and a technique for taking it into account is required. Surveys during the past twenty-five years have shown that, during this period, in the aggregate those persons who favor the Democratic Party or candidate in national elections are less likely to vote than those who favor the Republican Party or candidate. Even if this were not true, however, so long as only a part of the electorate votes, steps have to be taken to provide a sample of that particular part of the population, because it may well differ from the total population old enough to vote. The turnout ratio is particularly low in national nonpresidential elections. The ratios for the past five elections are shown in Table 2.

TABLE 2

PER CENT OF UNITED STATES POPULATION OLD ENOUGH TO VOTE WHO VOTED IN 1950, 1952, 1954, 1956, 1958

Election	*Type of Election*	*Per Cent Who Voted*
1950	Congressional	41.9
1952	Presidential	62.5
1954	Congressional	42.4
1956	Presidential	60.4
1958	Congressional	43.9

To make allowance for this factor we use questions such as these:

> How much thought have you given to the coming November elections—quite a lot, or only a little?
>
> Have you ever voted in this precinct or district?

Where do people who live in this neighborhood go to vote?

Are you now registered so that you can vote in the election this November?

Do you plan to register so that you can vote in the November election?

How often would you say you vote—always, nearly always, part of the time, or seldom?

Do you, yourself, plan to vote in the election this November?

How certain are you that you will vote—absolutely certain, fairly certain, or not certain?

In the election in November, 1956—when Eisenhower ran against Stevenson—did things come up which kept you from voting, or did you happen to vote?

Post-election studies indicate that the answers to these questions are related to voting participation, and indicate the extent to which they are related. A scoring system based on such studies is applied to the answers of respondents to each question. The scoring system is such that the greater their likelihood of voting, the higher their score. Respondents are then ranked on the basis of their scores.

Before using the turnout scale scores, all persons are excluded who say they are not registered and do not plan to register, or who say they do not plan to vote. Thus a person who is not registered and does not plan to register or who says he or she does not plan to vote is not included in the final sample of persons upon which the election estimate is based. These people are excluded because our studies and the studies of others have indicated that a negligible percentage of them vote, something on the order of between 1 and 5 per cent. Using the remaining sample of respondents, each class—on the basis of turnout scale scores—is tabulated on party or candidate preference (depending upon whether it is a national presidential or congressional election); the division of preference is then computed for the first scale position (those most likely to vote); the first and second scale positions combined; the first, second, and third scale positions combined; and so on. With this classification and accumulation, and the turnout ratio previously computed, that portion of the total sample with the greatest likelihood of voting (on the basis of these criteria) is selected, and the preference computed for that group. Usually this requires interpolation between classes.

THE UNDECIDED

For accurate estimation of the probable division of the vote, it is essential to minimize the proportion of replies which are in the category of undecided. For analytical purposes it is, of course, quite useful to use methods for maximizing the undecided, and this can be done quite easily. The opposite objective of reducing the undecided can be quite difficult, however, and, of course, when it is truly a case of being completely undecided, impossible. Two basic approaches we have adopted are (1) to use a question which makes the decision as easy as possible and (2) to ask the undecided person if he can say in which direction he is leaning in his preference. The question combination being used in current Gallup Poll surveys is:

a. If the Presidential election were being held TODAY, which candidates would you vote for—Nixon and Lodge or Kennedy and Johnson?
IF UNDECIDED OR REFUSED ASK:
  b. As of today, do you lean more to Nixon and Lodge or more to Kennedy and Johnson?

Before the "a" question is asked, the respondent is handed a card on which are printed the names of the two major parties and the respective candidates underneath their party designation.

The question has been framed to ask for the respondent's choice if the election were to be held "today" because it provokes less indecision than do questions which ask how a respondent expects to vote in November. Currently the use of the "leaning" question reduces the proportion of undecided by half. Sometimes the effect of a further reduction of the undecided can be obtained by the use of a secret ballot; this will be touched upon later.

COMPARING SURVEY RESULTS AND ELECTION RESULTS

Three sets of party- or candidate-preference survey results have been computed for each of the past five national elections, and the error computed for each. The three sets are (1) the total sample, with a shift adjustment based on the precinct vote in a previous election; (2) those in the sample who are registered or who plan to register and plan to vote, with shift adjustment; (3) the accumulated-classes method described above, with shift adjustment. The results are shown in Tables 3, 4, and 5, and provide a summation based on the final survey data illustrating how each step in the estimation procedure affected the results. As stated earlier, this summation is based on the vote excluding the South.

It should be emphasized that in actual practice we are unable to make complete use of the data obtained in the final pre-election survey. Insufficient time elapses between the last day of interviewing and the deadline for the release of the results to subscribing newspapers to receive the completed questionnaires through the mail, process them, and do a full analysis. In practice, in the case of the final survey, we have a simple basis for sorting out likely voters for the interviewers to use in their tabulations. The procedures described are used in processing the survey in early October; to the division of the vote obtained is applied any change observed between this survey and the final one.

TABLE 3

Comparison of Election Results and Survey Results in the Total Sample*: Per Cent Republican of the Democratic-Republican Vote

Election	(*N*)	*Survey Results*	*Election Results*	*Deviation*
1950	(2,452)	48.0	52.7	−4.7
1952	(4,814)	53.2	56.8	−3.6
1954	(4,784)	45.6	50.5	−4.9
1956	(6,173)	61.8	59.0	+2.8
1958	(4,625)	42.6	46.1	−3.5
Mean†				3.9
Range†				2.8-4.9

* Corrected for bias in terms of a previous election.
† Disregarding signs.

TABLE 4

Comparison of Election Results and Survey Results among Those Registered, Plan to Register, and Plan to Vote*: Per Cent Republican of the Democratic-Republican Vote

Election	(*N*)	*Survey Results*	*Election Results*	*Deviation*
1950	(1,999)	49.2	52.7	−3.5
1952	(4,229)	54.2	56.8	−2.6
1954	(3,733)	47.3	50.5	−3.2
1956	(5,438)	61.6	59.0	+2.6
1958	(3,662)	43.8	46.1	−2.3
Mean†				2.8
Range†				2.3-3.5

* Corrected for bias in terms of a previous election.
† Disregarding signs.

TABLE 5

Comparison of Election Results and Survey Results among Those Whose Turnout Score Placed Them in That Part of the Sample Corresponding to the Expected Turnout Proportion: Per Cent Republican of the Democratic-Republican Vote

Election	(*N*)	*Survey Results*	*Election Results*	*Deviation*
1950	(1,235)	52.4	52.7	−0.3
1952	(3,240)	55.3	56.8	−1.5
1954	(2,410)	50.0	50.5	−0.5
1956	(4,150)	62.1	59.0	+3.1
1958	(2,465)	45.9	46.1	−0.2
Mean*				1.1
Range*				0.2-3.1

* Disregarding signs.

In comparing the two surveys, of course, the simplified basis used by interviewers in separating likely voters from nonvoters for their telegraphic report is used for both. In the description here, however, the final survey materials as analyzed after the election have been used, because an accurate appraisal of the results of the methods can best be made with survey data obtained as close as possible to the election. The final survey interviewing period in the five elections covered in this description has centered five days before election day.

USE OF A SECRET BALLOT

Obtaining unbiased expressions of candidate preference is, of course, a matter of continuing great concern in election surveys. The traditional and obvious method of dealing with the problem is to use a secret ballot. Doing so in a completely satisfactory way in the typical survey situation presents difficulties, although the Gallup Poll has used a secret ballot at one time or another in election surveys from 1936 onward. Early in 1956 a method was worked out to incorporate a secret ballot procedure, making use of an approach developed by Sidney Goldish in the election surveys of the Minnesota Poll. In this approach, the interview is conducted in the usual manner (i.e. as an "open" interview) for all questionnaire items except the voting preference question. This enables the interviewer to establish rapport with the respondent in the manner customary in an "open" interview. After asking a series of introductory questions, the respondent is handed a card showing the parties and candidates and asked to check his choice privately, fold the card, and drop it in a ballot box carried by the interviewer. The remainder of the interview is carried out in the usual manner.

The secret ballot was employed in two Gallup Poll national surveys conducted in early October and mid-October of 1956 and in mid-September and mid-October of 1958. In these surveys an interpenetrating sample was used, with every second respondent interviewed using the secret questionnaire. A comparison of the two sets of data in 1956 showed a 60 per cent lower undecided in the secret survey and a relatively greater vote for the Democratic ticket of Stevenson and Kefauver than was true in the nonsecret survey. In 1958 the secret survey produced 30 per cent less undecided than the nonsecret survey and a relatively lower Democratic vote. An application of the secret-nonsecret differentials in 1956 lowered the Eisenhower-Nixon percentage of the vote one and a half percentage points in the non-South; in 1958 it produced a change of less than a percentage point.

Table 6 recapitulates the final estimates that were obtained using the final survey data in the non-South and utilizing all steps in the estimating procedure, including the secret ballot in 1956 and 1958. Actually, in 1956 and 1958 the difference between the secret and non-secret was statistically significant only in the case of the per cent undecided. Nonetheless, the differential was applied to the party vote for the final estimate in both years and improved matters in 1956.

TABLE 6

COMPARISON OF ELECTION RESULTS AND FINAL SURVEY RESULTS: PER CENT REPUBLICAN OF THE DEMOCRATIC-REPUBLICAN VOTE

Election	*Survey Results*	*Election Results*	*Deviation*
1950	52.4	52.7	−0.3
1952	55.3	56.8	−1.5
1954	50.0	50.5	−0.5
1956	60.6	59.0	+1.6
1958	46.3	46.1	+0.2
Mean*			0.8
Range*			0.2-1.5

*Disregarding signs.

CHAPTER FOUR

CLASS NOTES

The concept of random sampling is discussed at length in the selection of readings. Our purpose here is to provide some data in a convenient form so that the student can apply what he has learned--and, we hope, in the process learn it more thoroughly.

A table of random numbers consists of a large number of digits jumbled together in a chance fashion. (They are arranged in blocks of five merely to facilitate reading, in Table I, taken from George W. Snedecor, Statistical Methods.) There are 50 rows and 50 columns, each numbered from 00 to 49.

Since the numbers are arranged in a random fashion, any sequence of numbers can be considered random. You can begin at the upper left corner and read across, or read down; you can begin at any randomly selected point; and so on.

For example, to draw a random sample of 50 cases out of a list of 99 cases numbered one to 99, we can start at any point--let us say it is row 11, column 8. We then start by taking two digit numbers, reading from left to right:

04, 05, 33, 64, 71, 72, 64, 56, 90, 66, 33, 86, 67, 92, 39, and so on.

If any number is repeated, we simply discard it; and we continue the process until we have accumulated the 50 cases to be in our sample.

We have included a roster of households in a hypothetical town, giving the address, the race of the head of the household, the number of years of schooling of the head of the household, and the annual income of the head of the household. The student can number each of the addresses in the town, and then select a random sample of cases--40, let us say. He can then calculate the characteristics of the households in this sample (race, schooling, and income) and compare these to the distribution of characteristics in the town as a whole (the data are given at the end of the roster of households).

This use of a table of random numbers depends, of course, on having a convenient list of all the units in the population to be sampled. In many instances, unfortunately, such a list does not exist or would be too costly, or too time-consuming to construct. In this event, we must find other means to select a random sample.

One such method can be illustrated by using Figure 1. We can assume for present purposes that no accurate list of all the households in the town is available. Our procedure then might be as follows:

(a) We can first select a random sample of blocks--numbering each block, and using our table of random numbers to select a sample of blocks.

(b) We can then pick a starting point in each block at random. (If we start at the corner of each block, we may bias our sample. Corner lots tend to be more expensive.)

(c) We can then select every other house in the block, every third house, etc., depending on the size of the sample desired.

NOTE: We might try to list all the households in each block in our sample of blocks, assign a number to each household, and then select a random sample. This would be preferable, but takes a good deal more time.

What happens as we decrease the number of blocks in our sample of blocks and increase the number of households selected in each block, to get the same sample size?

What happens as we increase the number of blocks in our sample of blocks and decrease the number of households selected in each block, to get the same sample size?

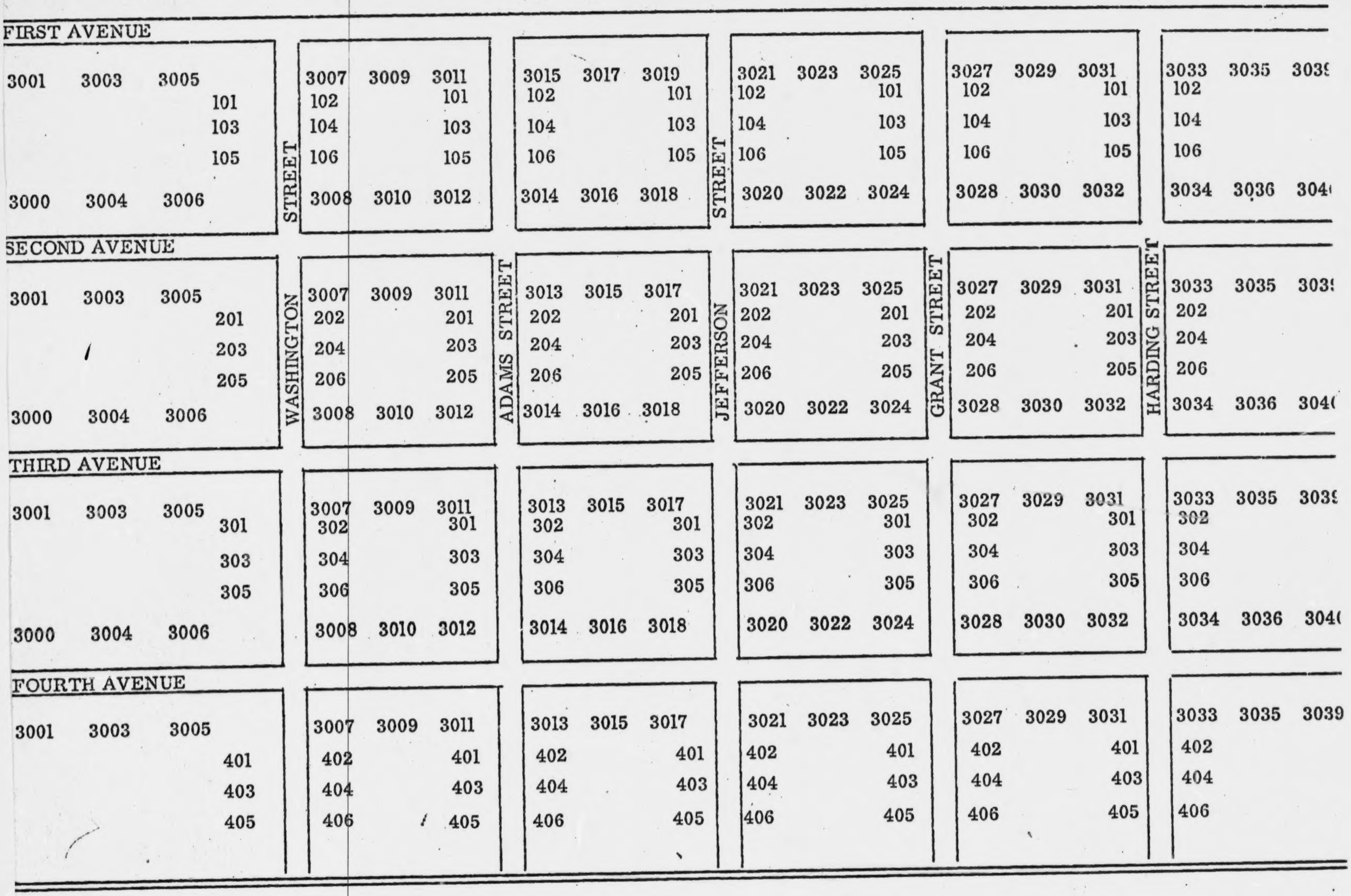
What happens if each block in the town has an equal chance of appearing in our random sample of blocks and if we take the same number of households from each block in the sample? Will each household in the population of households have an equal chance of appearing in the final sample?

TABLE 1.2

TEN THOUSAND RANDOMLY ASSORTED DIGITS

	00–04	05–09	10–14	15–19	20–24	25–29	30–34	35–39	40–44	45–49
00	54463	22662	65905	70639	79365	67382	29085	69831	47058	08186
01	15389	85205	18850	39226	42249	90669	96325	23248	60933	26927
02	85941	40756	82414	02015	13858	78030	16269	65978	01385	15345
03	61149	69440	11286	88218	58925	03638	52862	62733	33451	77455
04	05219	81619	10651	67079	92511	59888	84502	72095	83463	75577
05	41417	98326	87719	92294	46614	50948	64886	20002	97365	30976
06	28357	94070	20652	35774	16249	75019	21145	05217	47286	76305
07	17783	00015	10806	83091	91530	36466	39981	62481	49177	75779
08	40950	84820	29881	85966	62800	70326	84740	62660	77379	90279
09	82995	64157	66164	41180	10089	41757	78258	96488	88629	37231
10	96754	17676	55659	44105	47361	34833	86679	23930	53249	27083
11	34357	88040	53364	71726	45690	66334	60332	22554	90600	71113
12	06318	37403	49927	57715	50423	67372	63116	48888	21505	80182
13	62111	52820	07243	79931	89292	84767	85693	73947	22278	11551
14	47534	09243	67879	00544	23410	12740	02540	54440	32949	13491
15	98614	75993	84460	62846	59844	14922	48730	73443	48167	34770
16	24856	03648	44898	09351	98795	18644	39765	71058	90368	44104
17	96887	12479	80621	66223	86085	78285	02432	53342	42846	94771
18	90801	21472	42815	77408	37390	76766	52615	32141	30268	18106
19	55165	77312	83666	36028	28420	70219	81369	41943	47366	41067
20	75884	12952	84318	95108	72305	64620	91318	89872	45375	85436
21	16777	37116	58350	42958	21460	43910	01175	87894	81378	10620
22	46230	43877	80207	88877	89380	32992	91380	03164	98656	59337
23	42902	66892	46134	01432	94710	23474	20423	60137	60609	13119
24	81007	00333	39693	28039	10154	95425	39220	19774	31782	49037
25	68089	01122	51111	72373	06902	74373	96199	97017	41273	21546
26	20411	67081	89950	16944	93054	87687	96693	87236	77054	33848
27	58212	13160	06468	15718	82627	76999	05999	58680	96739	63700
28	70577	42866	24969	61210	76046	67699	42054	12696	93758	03283
29	94522	74358	71659	62038	79643	79169	44741	05437	39038	13163
30	42626	86819	85651	88678	17401	03252	99547	32404	17918	62880
31	16051	33763	57194	16752	54450	19031	58580	47629	54132	60631
32	08244	27647	33851	44705	94211	46716	11738	55784	95374	72655
33	59497	04392	09419	89964	51211	04894	72882	17805	21896	83864
34	97155	13428	40293	09985	58434	01412	69124	82171	59058	82859
35	98409	66162	95763	47420	20792	61527	20441	39435	11859	41567
36	45476	84882	65109	96597	25930	66790	65706	61203	53634	22557
37	89300	69700	50741	30329	11658	23166	05400	66669	48708	03887
38	50051	95137	91631	66315	91428	12275	24816	68091	71710	33258
39	31753	85178	31310	89642	98364	02306	24617	09609	83942	22716
40	79152	53829	77250	20190	56535	18760	69942	77448	33278	48805
41	44560	38750	83635	56540	64900	42912	13953	79149	18710	68618
42	68328	83378	63369	71381	39564	05615	42451	64559	97501	65747
43	46939	38689	58625	08342	30459	85863	20781	09284	26333	91777
44	83544	86141	15707	96256	23068	13782	08467	89469	93842	55349
45	91621	00881	04900	54224	46177	55309	17852	27491	89415	23466
46	91896	67126	04151	03795	59077	11848	12630	98375	52068	60142
47	55751	62515	21108	80830	02263	29303	37204	96926	30506	09808
48	85156	87689	95493	88842	00664	55017	55539	17771	69448	87530
49	07521	56898	12236	60277	39102	62315	12239	07105	11844	01117

FIGURE 1


FIRST AVENUE
3001 3003 3005
101
103
105
3000 3004 3006
STREET
3007 3009 3011
102 101
104 103
106 105
3008 3010 3012
3015 3017 3019
102 101
104 103
106 105
3014 3016 3018
STREET
3021 3023 3025
102 101
104 103
106 105
3020 3022 3024
3027 3029 3031
102 101
104 103
106 105
3028 3030 3032
3033 3035 303
102
104
106
3034 3036 304
SECOND AVENUE
3001 3003 3005
201
203
205
3000 3004 3006
WASHINGTON
3007 3009 3011
202 201
204 203
206 205
3008 3010 3012
ADAMS STREET
3013 3015 3017
202 201
204 203
206 205
3014 3016 3018
JEFFERSON
3021 3023 3025
202 201
204 203
206 205
3020 3022 3024
GRANT STREET
3027 3029 3031
202 201
204 203
206 205
3028 3030 3032
HARDING STREET
3033 3035 303
202
204
206
3034 3036 304
THIRD AVENUE
3001 3003 3005
301
303
305
3000 3004 3006
3007 3009 3011
302 301
304 303
306 305
3008 3010 3012
3013 3015 3017
302 301
304 303
306 305
3014 3016 3018
3021 3023 3025
302 301
304 303
306 305
3020 3022 3024
3027 3029 3031
302 301
304 303
306 305
3028 3030 3032
3033 3035 303
302
304
306
3034 3036 304
FOURTH AVENUE
3001 3003 3005
401
403
405
3007 3009 3011
402 401
404 403
406 405
3013 3015 3017
402 401
404 403
406 405
3021 3023 3025
402 401
404 403
406 405
3027 3029 3031
402 401
404 403
406 405
3033 3035 3039
402
404
406

ROSTER OF CITY HOUSEHOLDS

LOCATION	RACE OF HEAD OF HOUSEHOLD (Male)	NUMBER OF YEARS OF SCHOOL OF H OF H	ANNUAL INCOME OF H OF H
First Avenue			
3001	White	17	$ 15,000
3003	White	18	$ 16,000
3005	Negro	18	$ 14,000
3007	White	17	$ 15,000
3009	White	18	$ 14,000
3011	White	18	$ 14,000
3015	White	18	$ 15,000
3017	White	17	$ 14,000
3019	White	18	$ 14,000
3021	White	18	$ 13,000
3023	White	18	$ 12,000
3025	White	18	$ 13,000
3027	White	16	$ 8,000
3029	White	14	$ 8,000
3031	White	15	$ 8,000
3033	White	14	$ 7,000
3035	White	9	$ 7,000
3039	White	9	$ 7,000

LOCATION	RACE OF HEAD OF HOUSEHOLD (Male)	NUMBER OF YEARS OF SCHOOL OF H OF H	ANNUAL INCOME OF H OF H
Second Avenue			
3000	White	18	$ 16,000
3001	White	17	$ 15,000
3003	White	18	$ 16,000
3004	White	8	$ 11,000
3005	White	17	$ 16,000
3006	White	10	$ 15,000
3007	White	11	$ 14,000
3008	Negro	17	$ 12,000
3009	White	12	$ 11,000
3010	White	12	$ 11,000
3011	White	17	$ 16,000
3012	White	13	$ 11,000
3013	White	14	$ 9,000
3014	Negro	11	$ 10,000
3015	White	15	$ 9,000
3016	White	10	$ 8,000
3017	White	11	$ 8,000
3018	White	12	$ 8,000
3020	White	13	$ 8,000
3021	White	12	$ 7,000
3022	White	12	$ 7,000
3023	White	11	$ 6,000

LOCATION	RACE OF HEAD OF HOUSEHOLD (Male)	NUMBER OF YEARS OF SCHOOL OF H OF H	ANNUAL INCOME OF H OF H
Second Avenue			
3024	White	11	$ 6,000
3025	White	11	$ 6,000
3027	White	12	$ 6,000
3028	White	8	$ 5,000
3029	Negro	7	$ 8,000
3030	Negro	8	$ 5,000
3031	Negro	9	$ 6,000
3032	Negro	10	$ 6,000
3033	Negro	10	$ 6,000
3034	Negro	11	$ 5,000
3035	Negro	12	$ 5,000
3036	Negro	9	$ 6,000
3039	Negro	9	$ 5,000
3040	Negro	12	$ 8,000

LOCATION	RACE OF HEAD OF HOUSEHOLD (Male)	NUMBER OF YEARS OF SCHOOL OF H OF H	ANNUAL INCOME OF H OF H
Third Avenue 3000	White	17	$ 16,000
3001	White	17	$ 16,000
3003	White	15	$ 13,000
3004	Negro	16	$ 11,000
3005	White	11	$ 8,000
3006	White	13	$ 12,000
3007	White	11	$ 9,000
3008	White	12	$ 9,000
3009	White	13	$ 8,000
3010	White	11	$ 9,000
3011	White	11	$ 9,000
3012	White	11	$ 7,000
3013	White	12	$ 9,000
3014	White	10	$ 6,000
3015	White	13	$ 9,000
3016	White	14	$ 9,000
3017	White	15	$ 6,000
3018	White	10	$ 6,000
3020	Negro	13	$ 8,000
3021	White	8	$ 6,000
3022	White	9	$ 6,000
3023	White	8	$ 6,000

LOCATION	RACE OF HEAD OF HOUSEHOLD (Male)	NUMBER OF YEARS OF SCHOOL OF H OF H	ANNUAL INCOME OF H OF H
Third Avenue 3024	White	9	$ 5,000
3025	White	10	$ 5,000
3027	White	11	$ 5,000
3028	White	11	$ 5,000
3029	White	11	$ 4,000
3030	White	12	$ 4,000
3031	White	13	$ 4,000
3032	White	13	$ 3,000
3033	White	9	$ 3,000
3034	White	8	$ 3,000
3035	White	7	$ 2,000
3036	White	9	$ 2,000
3039	White	7	$ 3,000
3040	White	7	$ 3,000

LOCATION	RACE OF HEAD OF HOUSEHOLD (Male)	NUMBER OF YEARS OF SCHOOL OF H OF H	ANNUAL INCOME OF H OF H
Fourth Avenue 3000	White	13	$ 7,000
3001	White	13	$ 7,000
3003	White	14	$ 7,000
3004	White	10	$ 6,000
3005	White	11	$ 7,000
3006	White	12	$ 7,000
3007	White	10	$ 7,000
3008	White	10	$ 7,000
3009	White	9	$ 6,000
3010	White	10	$ 6,000
3011	White	10	$ 7,000
3012	White	10	$ 7,000
3013	White	10	$ 6,000
3014	White	11	$ 7,000
3015	White	10	$ 6,000
3016	White	9	$ 5,000
3017	White	10	$ 5,000
3018	White	9	$ 4,000
3020	Negro	15	$ 9,000
3021	Negro	9	$ 4,000
3022	Negro	10	$ 9,000
3023	Negro	11	$ 7,000

LOCATION	RACE OF HEAD OF HOUSEHOLD (Male)	NUMBER OF YEARS OF SCHOOL OF H OF H	ANNUAL INCOME OF H OF H
Fourth Avenue			
3024	Negro	11	$ 7,000
3025	Negro	11	$ 7,000
3027	Negro	8	$ 4,000
3028	Negro	9	$ 6,000
3029	Negro	8	$ 4,000
3030	Negro	10	$ 6,000
3031	Negro	8	$ 4,000
3032	Negro	8	$ 4,000
3033	Negro	7	$ 3,000
3034	Negro	6	$ 2,000
3035	Negro	6	$ 2,000
3036	Negro	7	$ 3,000
3039	Negro	7	$ 3,000
3040	Negro	6	$ 2,000

LOCATION	RACE OF HEAD OF HOUSEHOLD (Male)	NUMBER OF YEARS OF SCHOOL OF H OF H	ANNUAL INCOME OF H OF H
Washington Street			
101	White	16	$ 15,000
102	White	17	$ 14,000
103	White	17	$ 14,000
104	White	16	$ 13,000
105	White	16	$ 13,000
106	White	15	$ 13,000
201	White	16	$ 15,000
202	White	14	$ 11,000
203	White	17	$ 16,000
204	White	14	$ 15,000
205	White	15	$ 12,000
206	White	16	$ 11,000
301	White	15	$ 10,000
302	White	11	$ 7,000
303	White	11	$ 9,000
304	White	12	$ 7,000
305	White	11	$ 9,000
306	White	10	$ 7,000
401	White	11	$ 7,000
402	White	11	$ 6,000
403	White	10	$ 7,000
404	White	10	$ 6,000
405	White	10	$ 6,000
406	White	10	$ 5,000

LOCATION	RACE OF HEAD OF HOUSEHOLD (Male)	NUMBER OF YEARS OF SCHOOL OF H OF H	ANNUAL INCOME OF H OF H
Adams Street			
101	White	11	$ 13,000
102	White	12	$ 13,000
103	White	13	$ 12,000
104	White	14	$ 12,000
105	White	11	$ 11,000
106	White	11	$ 10,000
201	White	11	$ 10,000
202	White	13	$ 10,000
203	White	13	$ 10,000
204	White	11	$ 8,000
205	White	12	$ 10,000
206	White	11	$ 7,000
301	White	11	$ 7,000
302	White	11	$ 6,000
303	White	11	$ 6,000
304	White	11	$ 7,000
305	White	11	$ 6,000
306	Negro	14	$ 8,000
401	White	11	$ 6,000
402	White	12	$ 5,000
403	White	12	$ 5,000
404	White	12	$ 6,000
405	White	11	$ 5,000
406	White	11	$ 5,000

LOCATION	RACE OF HEAD OF HOUSEHOLD (Male)	NUMBER OF YEARS OF SCHOOL OF H OF H	ANNUAL INCOME OF H OF H
Jefferson Street			
101	White	16	$ 12,000
102	White	15	$ 11,000
103	White	16	$ 8,000
104	White	13	$ 10,000
105	White	11	$ 6,000
106	White	14	$ 10,000
201	White	11	$ 7,000
202	White	11	$ 6,000
203	White	11	$ 6,000
204	White	11	$ 5,000
205	White	11	$ 5,000
206	White	12	$ 7,000
301	White	12	$ 5,000
302	White	13	$ 5,000
303	White	12	$ 5,000
304	White	11	$ 4,000
305	White	12	$ 4,000
306	White	12	$ 4,000
401	Negro	10	$ 6,000
402	Negro	11	$ 6,000
403	Negro	12	$ 6,000
404	White	12	$ 4,000
405	White	12	$ 4,000
406	Negro	10	$ 6,000

LOCATION	RACE OF HEAD OF HOUSEHOLD (Male)	NUMBER OF YEARS OF SCHOOL OF H OF H	ANNUAL INCOME OF H OF H
Grant Street			
101	White	12	$ 12,000
102	White	14	$ 7,000
103	White	12	$ 10,000
104	White	13	$ 7,000
105	White	15	$ 6,000
106	White	15	$ 6,000
201	White	15	$ 5,000
202	White	15	$ 6,000
203	White	14	$ 5,000
204	White	13	$ 3,000
205	White	13	$ 5,000
206	White	14	$ 5,000
301	White	13	$ 4,000
302	White	13	$ 4,000
303	White	13	$ 4,000
304	White	13	$ 3,000
305	Negro	12	$ 4,000
306	White	11	$ 7,000
401	Negro	10	$ 5,000
402	Negro	12	$ 4,000
403	Negro	12	$ 4,000
404	Negro	9	$ 4,000
405	Negro	7	$ 4,000
406	Negro	8	$ 3,000

LOCATION	RACE OF HEAD OF HOUSEHOLD (Male)	NUMBER OF YEARS OF SCHOOL OF H OF H	ANNUAL INCOME OF H OF H
Harding Street			
101	White	14	$ 7,000
102	White	14	$ 7,000
103	White	14	$ 6,000
104	White	16	$ 6,000
105	White	14	$ 5,000
106	White	14	$ 6,000
201	White	14	$ 5,000
202	White	14	$ 5,000
203	White	15	$ 5,000
204	White	12	$ 4,000
205	White	9	$ 4,000
206	White	6	$ 4,000
301	Negro	8	$ 6,000
302	Negro	6	$ 3,000
303	White	8	$ 5,000
304	Negro	6	$ 4,000
305	Negro	8	$ 5,000
306	Negro	7	$ 4,000
401	Negro	6	$ 3,000
402	Negro	7	$ 2,000
403	Negro	8	$ 2,000
404	Negro	8	$ 3,000
405	Negro	6	$ 3,000
406	Negro	6	$ 2,000

Total Income	$ 1,838,000
Total Number of Households	246
Per Cent Negro Households	22.4
Mean Income	$ 7,470
Mean Years of Schooling	10.7

CHAPTER V

THE PROBLEM OF INFERENCE

I

In the previous chapter it was pointed out that we are frequently faced with the necessity of making estimates or statements based on a sample drawn from a larger population or universe. It is the technique of making these estimates, commonly called the process of inference, which now concerns us.

For some law students, the concepts and the notational system and the computations we will be dealing with will be perfectly familiar. They are, in fact, relatively simple and straight forward and they require no more mathematical background than the ability to add, subtract, multiply, and divide.

Before we turn to the process of statistical inference, however, it is necessary for the student to understand a number of basic ideas such as mean, variance, normal distribution, and so on. Any one of the books on elementary statistics listed in Appendix A will provide a clear and concise explanation.

I I

Two of these ideas--that of the standard deviation and the normal curve--are really essential for understanding the process of statistical inference; and it is this process which is discussed by John Mueller and Karl Schuessler in the selection from Statistical Reasoning in Sociology.

The most crucial idea in this selection by Mueller and Schuessler involves the null hypothesis. This is an idea that will, in all probability, provide some difficulty to the law student--and for most students, for that matter, if our experience is any guide. The law student, however, has one advantage, for the concept of the null hypothesis will probably be familiar to him in another guise--namely, the judgment of guilty beyond a reasonable doubt as it appears in criminal trials.

Now many legal writers are agreed that the precise meaning of "beyond a reasonable doubt" is far from clear and that it is a notion which has confused many a jury. But one way to look at it, we think, is along the following line of reasoning. The defendant in a criminal trial is presumed "innocent until proven guilty." The prosecution then presents a body of evidence showing that the defendant is guilty of the crime with which he is charged--a body of evidence some part of which still stands after the defense attorney has held it up for criticism and presented his countervailing facts. Now the jury must ask itself: What are the chances that this array of evidence pointing to guilt could have arisen if in fact the defendant is innocent?

If it appears that the defendant could be innocent and yet there is still a good chance that such an array of evidence pointing to guilt could arise fairly often (50 out of 100 times, let us say) the jury may be very reluctant to bring in a verdict of guilty--the jury would rather cling to the assumption of innocence. Indeed, if such an array of evidence pointing to guilt would arise relatively rarely (only 20 out of 100 times, let us say) when the defendant was innocent, the jury still may not feel it has enough grounds to convict, i.e., reject the assumption of innocence. But if the chances are very slim indeed that an innocent man would every be confronted with such proof of guilt (once outof 100 times, perhaps) the jury may feel that it must reject its assumption of innocence and declare the defendant a criminal. The members of the jury have been convinced, we say, beyond a reasonable doubt.

In criminal trials, of course, there is almost never an opportunity to attach any precise figure to the chances of a body of evidence as a whole pointing to guilt occurring and the jury is forced to rely on more impressionistic evaluations. Nonetheless, the logical process of making a judgment about guilt in a criminal trial and testing a null hypothesis in statistics seem to show many similarities.

III

The last selection in this chapter deals with the application of sampling theory and statistical inference to the problem of discrimination in the selection of juries. It represents, we think an outstanding illustration of how the disciplines of law and the social sciences can and must work on certain problems arising in a large, complex society. Certainly the social sciences by themselves would find it difficult to see exactly how their knowledge of quantitative methods and patterns of discrimination could be made most relevant for the legal disputes about the composition of the jury. And there is little doubt but that the law alone would find problems in seeing both the possibilities and the limitations of ".01" levels, etc. When they join forces, however, they can exert great leverage in the maintenance of individual rights and the use of law for social ends. And the possibility of fruitful cooperation justifies, in our opinion, the exposure of the law student to materials which are likely to seem somewhat outside his regular course of study.

STATISTICAL BACKGROUND

Sampling, as we have seen, is a useful technique for obtaining information about the characteristics of a universe. The information derived from a sample, however, can only give an estimate of the characteristics present in the universe. The fact that a random procedure is used in drawing a sample helps to insure that the sample is not subject to systematic bias, but the results of any given sample are, in all probability, in error to some degree. That is, sampling does not give an exact picture of the characteristics under observation. For example, Mueller and Schuessler, in the following reading, present a universe of suicide rates in 107 American cities; the mean suicide rate of all the cities is 12.4. They then draw 100 random samples from the universe, four of which coincide with the true mean; and "approximately two-thirds of the 100 means fall between 11.4 and 13.3; or virtually within one point of the true mean." What is significant here is that sampling provides us with a fairly accurate estimation of the parameter, yet a large proportion of samples are in error. It is the procedure for dealing with such sampling errors that is discussed by Mueller and Schuessler.

From John H. Mueller and Karl F. Schuessler, Statistical Reasoning in Sociology, (Boston: Houghton Mifflin Company, 1961) pp. 361-379; 388-391. Reprinted by permission.

John H. Mueller and Karl F. Schuessler, Statistical Reasoning, pp 361-379; 388-391.

SECTION ONE

Parameter Estimation

Problems in Estimation. It should be unnecessary to repeat that a sample is useful only for the information it supplies on the characteristics of the universe. Thus, the sampling process finds its ultimate consummation in a description of the population from which the sample is drawn. But this description can be only an approximation: an average monthly expenditure of $92.75 in a random sample of college students will not correspond exactly to the average expenditure of the whole student body which the sample is designed to represent. Nor, obviously, will a second sample of $88.62 necessarily correspond more closely to the unknown parameter. If, therefore, owing to the vagaries of chance sampling, no two samples are alike and all are in error, with how much confidence can we speak of the value of the universe? Clearly, we cannot merely project the value of the sample onto the universe, and let it go at that. This procedure of adopting the sample value in lieu of the parameter, or universe value, is hedged about with as many regulations as is the procedure of sampling itself. We call this set of prescribed procedures *statistical inference*.

Since, in the practical affairs of life, it is so rare that we are able to examine a whole universe, and since sampling is so frequently the resort, the procedures of statistical inference loom quite large in the repertory of statisticians. In fact, some would actually identify the science of statistics with the problem of sampling and decision-making — an extreme emphasis which seems to the authors unrealistic, since descriptive statistics have a validity and importance in their own right.

The problems of statistical inference begin with the legitimate assumption of a discrepancy between the variable sample estimate and the universe parameter, which is constant. When this discrepancy is the result of random sampling, we call it a *sampling error*. If the average monthly expenditure of all students is $95.50 and the sample mean is $92.75, the sampling error is $2.75. In general, the sample rarely provides us with a perfect description of the universe, for the reason that few if any sample values are entirely free of sampling error.

Sampling Distribution. Even though such errors cling to almost every sample statistic, there is no reason to be dismayed by such a natural consequence of random sampling. These errors, when taken in the mass, do not behave chaotically; on the contrary, they behave systematically, they exhibit a characteristic distribution, and they are therefore conveniently amenable to analysis. There is not only a definite zone surrounding the parent value beyond which the variable samples cannot stray, but in addition the variety of sample values distribute themselves in a recognizable pattern which permits statistical inferences to be confidently drawn. We can, for example, be certain that the average age of a sample of high school students will not fall below 10, or exceed 21; and reasonably certain that it will not fall short of 14 or surpass 17. We can assert this because we are so thoroughly familiar with the ages of high school students. But whether or not we are intimately familiar with the universe, we can still be assured that the sample value will always bear some resemblance to the corresponding universe value. We can proceed with confidence that the statistic will always adhere, to a greater or lesser extent, to its parameter.

To bring this very important fact home to us, we shall illustratively take a known universe, extract a large number of samples from it, calculate the mean for each, and then compare these means with the population mean. We will then have demonstrated to what extent the sample means actually digress from, or converge on, the parameter itself. Of course, in an actual problem, we will not have knowledge of the universe mean, nor will we take a large number of samples. Instead we will be ignorant of the true mean and will take only one sample, which we will have to make do. But with the background afforded by this experiment, we will be fortified with knowledge of how samples in general behave, all of which will help us to infer how the particular sample, which we have drawn, is probably behaving.

For our working universe, we shall take the previously analyzed suicide rates of 107 American cities (Table 3.1.1a). From this universe we shall (1) draw 100 random samples — restoring each sample to the universe before drawing another — and (2) compute the mean of each sample. This experiment ought to be sufficient to reveal something about the regularity in the behavior of the seemingly wayward samples. These experimental means are arrayed in Table 12.1.1a and grouped in Table 12.1.1b.

Table 12.1.1a   *Array of 100 Sample Means, n = 30*

9.9	11.6	12.2	12.7	13.1
10.0	11.6	12.2	12.7	13.2
10.3	11.7	12.2	12.8	13.2
10.3	11.7	12.2	12.8	13.2
10.7	11.8	12.3	12.8	13.3
10.8	11.8	12.4	12.8	13.3
10.8	11.8	12.4	12.9	13.3
10.9	11.8	12.4	12.9	13.3
10.9	11.9	12.4	12.9	13.4
11.0	12.0	12.5	12.9	13.5
11.0	12.0	12.5	12.9	13.5
11.1	12.0	12.5	13.0	13.5
11.1	12.0	12.6	13.0	13.6
11.1	12.0	12.6	13.0	13.7
11.3	12.1	12.6	13.0	13.8
11.4	12.1	12.6	13.1	13.9
11.5	12.1	12.6	13.1	14.0
11.5	12.1	12.6	13.1	14.0
11.5	12.2	12.6	13.1	14.2
11.6	12.2	12.6	13.1	14.3

Table 12.1.1b   *Frequency Tally, Empirical Sampling Distribution, 100 Sample Means, n = 30*

CLASS INTERVAL	TALLY	*f*
9.0– 9.9	/	1
10.0–10.9	𝍸 ///	8
11.0–11.9	𝍸 𝍸 𝍸 𝍸	20
12.0–12.9	𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 //	42
13.0–13.9	𝍸 𝍸 𝍸 𝍸 𝍸	25
14.0–14.9	////	4
		100

Inspecting this lot, we see that only 4 of the 100 samples coincide exactly with the universe mean, known to be 12.4. All others contain a sampling error; in fact, every one of them would probably have shown a sampling error if we had insisted on greater decimal accuracy. We can see, however, that the universe mean is snugly nested in the very middle of the whole array of sample means, and that most of the sample means are compactly huddled around the true mean of 12.4, which is practically at the 50 per cent division point. Approximately two-thirds of the 100 means fall between 11.4 and 13.3, or virtually within one point of the true mean. Characteristically, the sample means gravitate toward the parent mean, or the center of the universe distribution.

It now becomes evident how improbable it is that a sample mean will deviate seriously from the universe mean. Furthermore, not only do sample means cluster around the true mean, but their pattern of distribution takes on a shape unmistakably approaching the normal curve (Figure 12.1.1). If we had drawn, processed, and tabulated all * possible samples, instead of merely 100 of them, we would have a *pattern* of distribution which would display still greater conformity to the smooth normal curve. This hypothetical frequency curve of all possible sample means is labeled the *theoretical sampling distribution*, and has, as its mathematical model, the ideal normal curve. Any experimental distribution of a limited num-

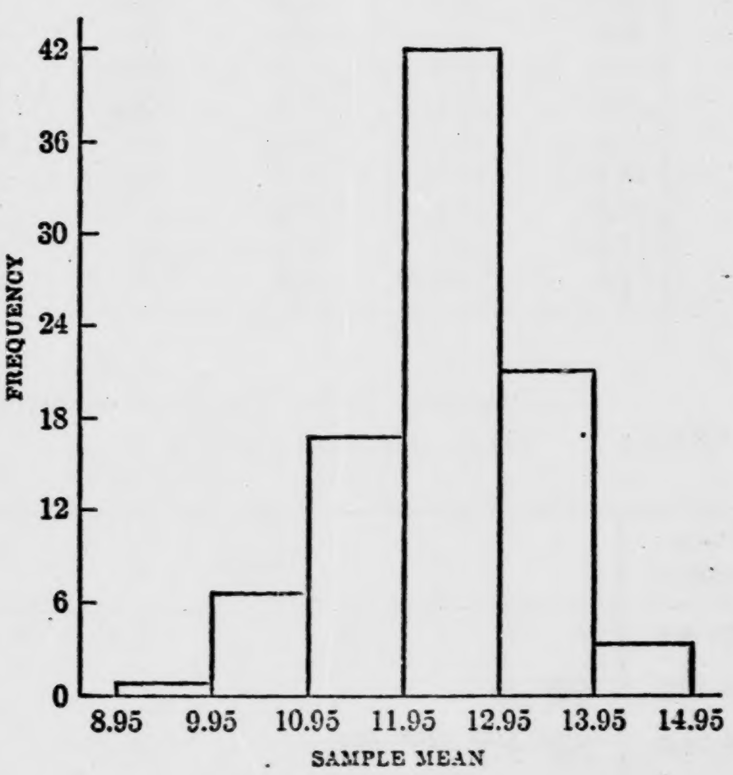


FIGURE 12.1.1 *Histogram of 100 Sample Means, n = 30*

ber of samples is called an *empirical sampling distribution*. The theoretical sampling distribution would, of course, be attained only by an infinite number of samples, since it would require that many to assure the distribution of sample means in their proper, expected proportions. The significant procedural point, however, in this type of problem, is that we can usually treat the theoretical sampling distribution of means as if it were a perfect, smooth, ideal normal curve.

---

* $C_n^N = C_{30}^{107} = \frac{P_{30}^{107}}{30!}$

Distribution of Sampling Errors. By subtracting the universe mean from each of the 100 sample means, we obtain the 100 sampling errors (Table 12.1.2a), which of course have the same curve pattern as the means themselves (Table 12.1.2b). All we have done is move the zero origin to the universe mean (12.4).

Table 12.1.2a *Array of Sampling Errors,* ***100*** *Samples,* $n = 80$

−2.5	−.8	−.2	+.3	+.7
2.4	.8	.2	.3	.8
2.1	.7	.2	.4	.8
2.1	.7	.2	.4	.8
1.7	.6	.1	.4	.9
1.6	.6	0	.4	.9
1.6	.6	0	.5	.9
1.5	.6	0	.5	.9
1.5	.5	0	.5	1.0
1.4	.4	+.1	.5	1.1
1.4	.4	.1	.5	1.1
1.3	.4	.1	.6	1.1
1.3	.4	.2	.6	1.2
1.3	.4	.2	.6	1.3
1.1	.3	.2	.6	1.4
1.0	.3	.2	.7	1.5
.9	.3	.2	.7	1.6
.9	.3	.2	.7	1.6
.9	.2	.2	.7	1.8
.8	.2	.2	.7	1.9

The distribution of sampling errors (Table 12.1.2b) again demonstrates that the small deviations are very numerous; the larger discrepancies are few in number. If we had drawn a single sample only, we could be practically certain that our sample mean would not have missed the true mean by more than two points, plus or minus; indeed, 68 per cent of the samples are in error by less than plus or minus one.

Table 12.1.2b

Frequency Distribution of Sampling Errors

CLASS INTERVAL	*f*
−3.4–(−2.5)	1
−2.4–(−1.5)	8
−1.4–(− .5)	20
− .4– .5	42
.6– 1.5	25
1.6– 2.5	4
	100

The presence of sampling errors obviously complicates the task of drawing an inference about the parameter; but the neat pattern of these errors renders the uncertainties of such inference less formidable. Our dissection of the empirical sampling distribution has given us considerable reassurance of the reasonableness of the behavior of samples, which will permit us to estimate the degree of confidence of which the single sample is worthy.

Sample Reliability. We must now take up the technique which will make appropriate allowance for the ever present sampling error, and thereby enable us to make reliable estimates of the parameter. For, unless we can rely with a certain degree of confidence on our inferences, there is no point in making them at all. The entire purpose of sampling is to substitute a sample of given reliability for the prohibitively complete enumeration of the unknown universe.

But we cannot measure the sampling error of our single sample directly since we are never given the true mean from which to compute it. That would suppose that we already have what we have set out to find in the first place. The single sample must ultimately therefore provide two kinds of information, as can be deduced from the foregoing analysis: (1) an estimate of the parameter, and (2) the degree of confidence we may place in that estimate, or the *reliability of the estimate* — in our example, the reliability of the mean.

The procedures for measuring the reliability of an estimate are numerous, since they differ according to the sample statistic (for example, whether mean or percentage) and according to the make-up of the sample (whether large or small, simple or stratified). Here, we present only two of the simplest techniques for measuring sample reliability; these apply to means and percentages of *large random samples*, respectively. They will suffice for our purposes, since other techniques produce results which in principle carry the same interpretation as the method here unfolded.

Interval Estimate of the Mean. Let us suppose that, in taking a single sample ($n = 30$) from the 107 suicide rates, we obtain a mean of 13.5. In such a realistic situation, we do not know the value of the true mean, and will never know it. At this stage of the inquiry, 13.5 is the only estimate we have of the true mean, and therefore it is our "best" estimate. We call this a *point estimate*. No one would claim that this estimate is any more than approximately correct. The weakness of the point estimate is that it provides no indication of the sampling error, and consequently we do not know how much confidence to accord it. This disadvantage is overcome by setting up an interval which probably contains the true mean, and which we therefore label an *interval estimate*. To construct such an interval, we (1) estimate the range within which a high proportion of sampling errors lie, and (2) add and subtract this range of error to and from our observed sample mean to form the interval which probably encloses the true mean.

This attack on the problem is analogous to the following homely situation. Encountering a man on the street, we may estimate his age to be 35 (point estimate). Upon being challenged, we reply: "Of course, this is only an estimate based on cues that I recognize. I feel reasonably confident, however, from samples of men I have known that he is somewhere between 30 and 40 (interval estimate); and feel still more confident that he is between 25 and 45; and absolutely certain that his age will be embraced between 20 and 50." These intervals enjoy varying degrees of confidence, and for that reason are more realistically known as *confidence intervals*. The larger the interval, the more confidence we have that it includes the true age. The age of the gentleman is, of course, fixed. It is what it is, even though we are not completely informed. But as we shift the size of the interval, contracting and expanding it around the point estimate, we revise the degree of our confidence that the interval will enclose his true age. To be sure, the initial estimate of 35 years may have been brilliantly correct, but we shall never have the satisfaction of knowing it unless we ask him directly — which under our assumptions we cannot do. So it is with sampling.

We already know how to draw a random sample; all we need to learn in addition is how to construct the aforementioned confidence interval, the first step of which is to establish the value of its unit of measure. This is the average by which all samples deviate from the unknown parameter, and is called the *standard error*.

The Logic of the Standard Error of the Mean. Our demonstration has revealed that the theoretical sampling distribution of means may be viewed as a normal curve (Figure 12.1.1), with the true mean at the center, as the grand mean of means. Statisticians express this truth more formally: "The sampling distribution of the mean is approximately normal around the true mean, provided $n > 30$." Now, we do not know whether the mean of the sample which we happened to draw lies near the true mean, a sigma removed from it, or at any other point on the base line; perhaps it is even three sigmas removed. But, if we had the value of the standard deviation of the sampling distribution, we could set up an interval estimate for the true mean; hence, we must obtain an estimate of this value. We label this value, *not* a standard deviation (*SD*), but more descriptively a *standard error* (*SE*), since the deviations of all sample means from the true mean are actually sampling errors.

Our sample alone must furnish the information on which to base an estimate of the standard error, since it is all we have. Hence, we must explore the sample to uncover some cues to the magnitude of the sought-for standard error. Initially, we might guess that the size of the sample (n) and the degree of scatter in the sample (*SD*) are related to the size of the standard error, according to the following reasoning.

First, as to sample size, we should expect larger samples to be more representative of the universe than smaller samples, and we would therefore expect larger samples to yield smaller sampling errors. In fact, when a sample contains every last item in a finite universe (a 100 per cent sample), there can be no sampling error at all. Every "sample" mean would then necessarily be identical with the universe mean, and the standard error would be zero. At the other extreme, the smallest possible sample would contain only one case ($n = 1$). The distribution of sample means would then be identical with that of individual universe values, and would display the same amount of variation as the items in the universe itself. Hence, in that instance, the standard error would be equal to the standard deviation of the universe (σ). From this brief analysis, we may conclude that the standard error of the mean (a) will always be smaller than the standard deviation of the universe, since n will always be larger than 1; and (b) will decrease as sample size (n) increases, reaching zero when $n = 100$ per cent of the universe.

Second, we intuitively expect the degree of scatter in the universe, as reflected by the degree of scatter in the sample, to affect the size of the sampling errors. Thus, if all values in the universe were alike, all sample means would also be alike, and there would be no sampling error whatever. The standard error of such a completely homogeneous universe would always be zero! A sample of one boy would give us without sampling error the average number of arms for all boys, since every boy has the same number of arms — a perfectly homogeneous universe; however, a sample of one boy would not give us without error the average height of all boys. There is a great variety of heights, which produces comparable variety within samples, and consequently among the sample means as well. Inverting this argument, we conjecture that a high degree of scatter in the sample is indicative of heterogeneity in the universe and therefore of large sampling errors. We thus anticipate that the standard error for a given n will be larger when sampling a very heterogeneous mass of data than when sampling a relatively homogeneous mass.

Computation of the Standard Error. Both of the foregoing insights are in accord with statistical theory, and are actually explicit in the formula for the standard error of the mean, which contains n and σ as its basic terms. Without further proof, we give the formula:

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} \tag{12.1.1}$$

where $\sigma_{\bar{X}}$ = standard error of the mean
σ = standard deviation of the universe
n = sample size

Of course, we do not have the standard deviation of the universe, as required by the formula; all we have is the standard deviation, s,* of the sample. We will therefore substitute it for the unknown parameter in the above formula.

$$s_{\bar{X}} = \frac{s}{\sqrt{n}} \tag{12.1.2}$$

But the value of the substituted standard deviation of the sample is not identical with that of the universe, owing to sampling error; in fact, the former tends to be smaller than the latter because the scatter of values within the sample tends to be smaller than that in the universe. To match this understatement in the numerator (i.e., to maintain the ratio in the fraction), we correspondingly reduce sample size (n) by one in the denominator:

$$s = \sqrt{\frac{\Sigma x^2}{n-1}} \tag{12.1.3}$$

This formula may fittingly be substituted for the population sigma. The correction becomes trivial, of course, if sample size is very large. Nevertheless, it is conventionally included, as a matter of principle, even in large samples. The working formula, with familiar symbols, will therefore read as follows:

$$s_{\bar{X}} = \frac{\sqrt{\dfrac{\Sigma x^2}{n-1}}}{\sqrt{n}}$$

$$= \sqrt{\frac{\Sigma x^2}{n(n-1)}} \tag{12.1.4}$$

This expression, then, yields the standard error of the mean which we have been looking for and from which we shall construct the interval estimate.

---

* Greek letters are conventionally used to represent *parameters* (e.g., σ, μ); Latin letters (e.g., s, $\bar{X}$) represent *sample statistics*, which are necessarily estimates. But this practice is not uniformly adhered to by writers in the field.

Construction of the Confidence Interval (*Large Sample*). Since the sampling distribution of the mean is approximately normal around the population mean, it follows that two out of three sample means (68.27 per cent) will lie within one standard error of the true mean. Our particular sample mean will therefore have a 2 in 3 chance of falling within one standard error of the true mean. Hence, by adding and subtracting one standard error to and from our randomly selected sample mean, we have an approximately 2 in 3 chance of enclosing the true mean. Similarly, 95 out of 100 sample means lie within 1.96 standard errors of the true mean; in consequence, if we attach 1.96 standard errors to either side of the sample mean, we have a 95 per cent probability of enclosing the true mean. In general, it is possible to provide any desired confidence interval by the simple technique of attaching the requisite multiple of the standard error to the observed sample mean. By thus making our interval estimate wide enough, we may be practically certain that the true mean has been contained within the interval, even though the observed sample mean is highly inaccurate.

Let us suppose that we have drawn a "bad" sample, whose mean, $\bar{X}_o$, is located in the tail of the sampling distribution, far removed from the universe mean (μ) which it intends to represent (Figure 12.1.2). Clearly, in this instance, it would not have been sufficient to attach only one standard error to either side of the observed mean in order to catch the true mean; nor would it have been enough even to attach two standard errors, since the true mean is more than two standard errors distant from our illustrative sample mean. However, if we affix three standard errors to either side of the observed mean, we obtain an interval whose lower limit extends substantially beyond the true mean. Since practically all (99.74 per cent) of the sample means lie within three standard errors of the universe mean, by adding and subtracting three standard errors to

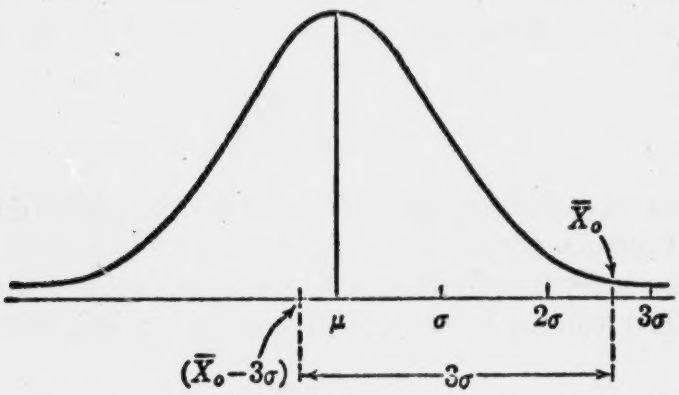


FIGURE 12.1.2 *Theoretical Sampling Distribution of Mean, Observed Sample Mean,* $\bar{X}_o$

and from any randomly selected mean we become virtually certain that the resultant interval will trap, as it were, the target mean. Even the most pessimistic soul, who always fears the worst, may give himself a sense of relative security by attaching three standard errors to his observed sample mean.

This method, be it noted, does not disclose the exact whereabouts of the true mean, as has been previously cautioned; it merely furnishes a stronger or weaker expectation that the true mean will be found within the specified interval estimate, derived from the observed sample. The true mean is stationary, wherever it is. Since it is the interval estimate that either succeeds or fails in enclosing the parameter on a given sampling trial, we may properly speak of the probability of such an interval succeeding in encompassing the universe mean. Of course, once the sample has been drawn and the interval formed — once the trial is over — it either does or does not include the target mean between its calculated limits, although we will never know. Our confidence rating in that interval corresponds to its pre-trial probability, which has been selected for our convenience and purpose.

Working Procedure. To illustrate the process of calculating a confidence interval, we carry out all necessary operations on the following sample of 30 suicide rates:

6.0	18.6	4.6	11.2	18.8
24.9	9.1	14.7	11.6	22.4
13.4	11.5	8.2	28.8	10.0
5.4	7.5	26.1	8.6	8.1
23.4	11.6	14.8	13.6	6.0
8.4	10.0	10.4	9.6	15.3

(1) We compute the sample mean:

$$\bar{X} = \frac{392.6}{30} = 13.1$$

(2) We compute the standard error of the sample mean:

$$s_{\bar{X}} = \sqrt{\frac{1{,}241.2}{(30)(29)}} = 1.2$$

(3) We multiply $s_{\bar{X}}$ by the number of sigma units to be added and subtracted for the agreed-upon confidence interval — for example, 1.96 for the 95 per cent interval:

$$1.96 s_{\bar{X}} = 1.96(1.2) = 2.4$$

(4) We add and subtract $1.96s_{\bar{X}}$ to and from the observed sample mean, $\bar{X}$. Thus:

$$\bar{X} \pm 1.96s_{\bar{X}}$$
$$13.1 + 2.3 = 15.4$$
$$13.1 - 2.3 = 10.8$$

(5) We finally make the interpretation that the chances are 95 in 100 that the true mean will be found within the interval 10.8–15.4. We are reasonably sure that our interval contains the true mean, since only 5 intervals in 100 constructed in the same manner will fail to do so. If, however, it doesn't, something has happened which would occur only 5 times in 100, a risk which we may be willing to face.

But how decide on the width of the confidence interval? Here one can offer only the most general guidance. Into the choice will enter a consideration of the consequences of the decision, the risk one will wish to run, and even the temperament of the person concerned. In effect, a decision like this is comparable to any other of the numerous decisions we make in the face of life's uncertainties.

Demonstration of Experimental Confidence Intervals. All statements of probability, it will be recalled, are based on the assumption of an infinite number of trials. While it is not possible to conduct such an infinity of trials, it is usually possible to conduct a fairly large number of experimental trials, the outcomes of which then serve to test the initial probability statement. To that end, we have calculated the 95 per cent confidence limits for each of our 100 samples ($n = 30$) of suicide rates, in order to determine whether the resulting intervals would actually enclose the true mean approximately 95 per cent of the times. The diagram on the opposite page (Figure 12.1.3) portrays the locations of the 100 confidence intervals, along with the known population mean of 12.4. It will be observed that exactly 95 out of 100 interval estimates do enclose the true mean, a result which is, surprisingly enough, identical with theoretical expectation. Here is tangible evidence that the confidence we place in a properly constructed interval estimate is amply justified.

Convenience of Large Samples. Up to now, all probability statements have been based on the assumption that the sampling distribution of the mean is normal. It is reassuring that such a sampling distribution will always be normal when the universe itself is normal. A normal population produces normality in this sampling distribution, whatever the size of the sample.

However, it is particularly characteristic of sociological data that populations are often not normal. Size of families, income and other social variables distinguish themselves from typical psychological data in that

John H. Mueller and Karl F. Schuessler

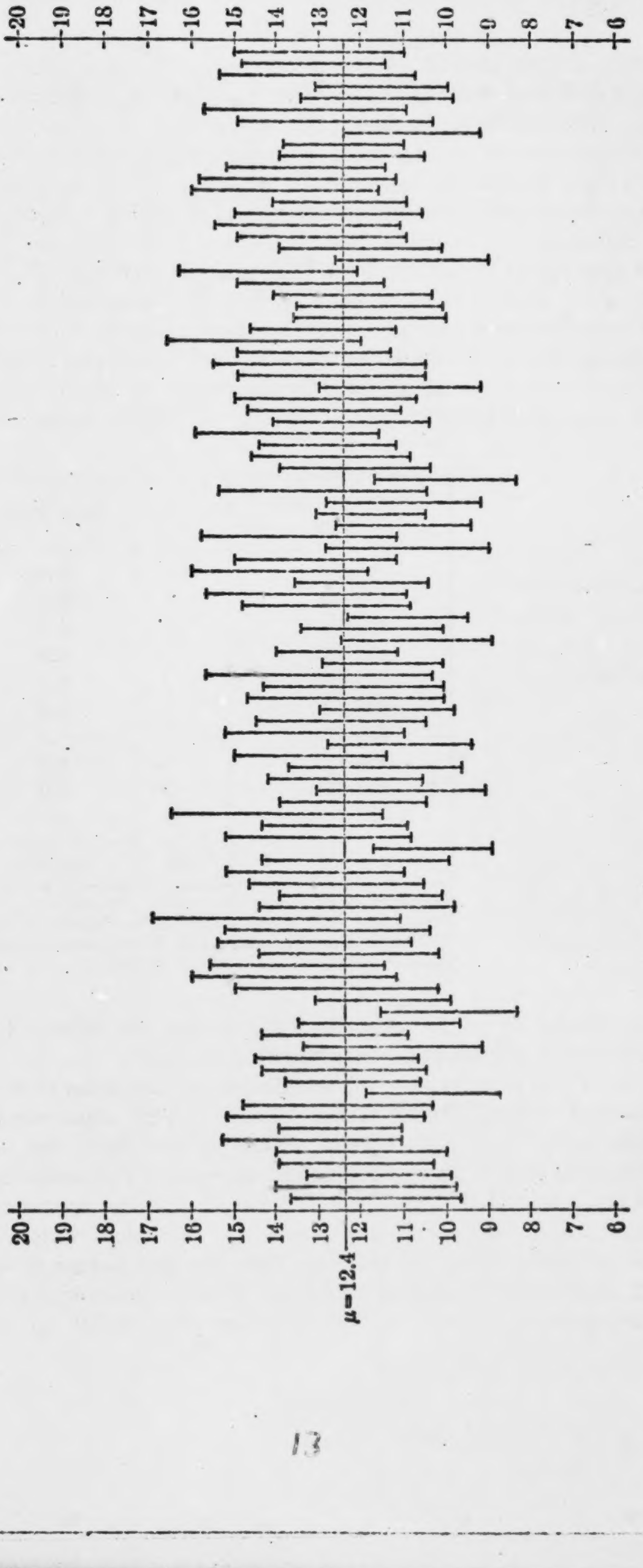


FIGURE 12.1.3 100 Interval Estimates, 95 Per Cent Confidence, n = 30; Universe: Suicide Rates, 107 U.S. Cities, 1950

they are often severely skewed. We may therefore raise the question of whether such skewness disturbs the required normality of the sampling distribution. The answer is: it does. But at the same time, the effect of such skewness can be circumvented by appropriately increasing the size of the sample beyond the approximate minimum of 30. As sample size increases, the sampling distribution approaches normality. However, no general statement can be made about the rate at which the sampling distribution approaches normality, since this rate will vary according to the severity of the skew to be offset. But we may gain some insight into what might on its face seem a rather remarkable phenomenon by examining the following experiment, in which we drew samples of size 20 and samples of size 40 from a compilation of 106 large American cities for size of population. The original distribution (Table 12.1.3) was highly skewed and consequently ideally suited for experimentally testing the effect of a skewed population on the sampling distribution of the mean.

Table **12.1.3**

Population Distribution, 106 American Cities 100,000 and Over, by Size of Population

SIZE OF CITIES ('000)	*f*	PER CENT
100–199	56	52.9%
200–299	14	13.2
300–399	11	10.4
400–499	7	6.6
500–599	5	4.7
600–699	2	1.9
700–799	1	.9
800–899	3	2.8
900–999	2	1.9
1,000 or more	5	4.7
	106	100.0%

Source: U.S. Bureau of the Census, *U.S. Census of the Population: 1950, Characteristics of the Population*, Vol. II, Part I, U.S. Summary, U.S. Government Printing Office, Washington, D.C., 1953, Table 1b.

The first set of 200 samples ($n = 20$) was drawn and the mean of each sample calculated and tabulated (Table 12.1.4a). This experimental sampling distribution is still perceptibly skewed to the right, but not nearly to the same degree as was the parent universe. This reduction in skewness is a result of the fundamental principle that the means are great levelers which necessarily cut down the individual extremes found in the universe. It follows therefore from the very nature of the mean that a distribution of sample means can never be as irregular, or as widely dispersed, as are the individual values which make up the

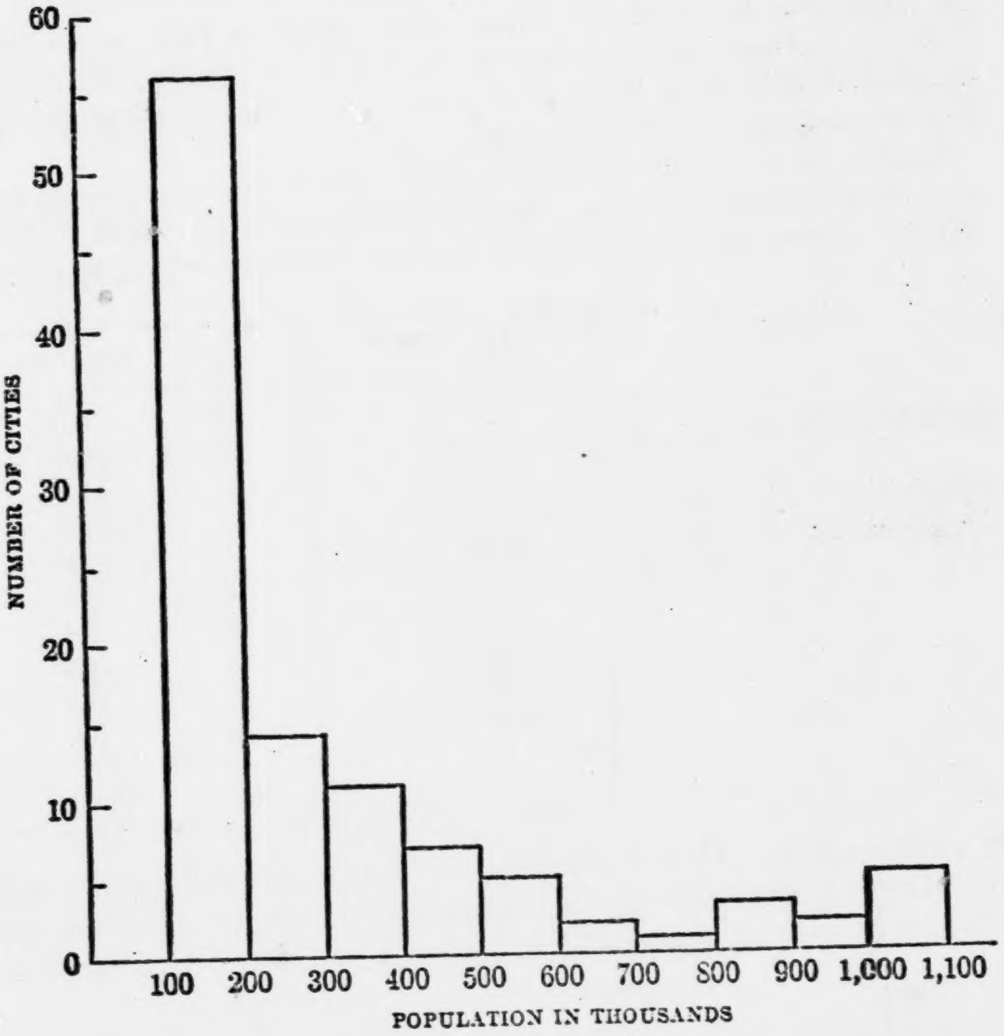


FIGURE 12.1.4 *Histogram, 106 American Cities by Size of Population*

Table 12.1.4a

Distribution of 200 Sample Means, n = 20

MEAN ('000)	f	PER CENT
100–199	2	1.0%
200–299	50	25.0
300–399	67	33.5
400–499	31	15.5
500–599	18	9.0
600–699	17	8.5
700–799	8	4.0
800–899	4	2.0
900–999	1	.5
1,000 or more	2	1.0
	200	100.0%

samples. Furthermore, the larger the sample, the more successfully will the dispersion be reduced. Thus, when sampling from a markedly skewed population, samples even as small as 20 will produce a sampling distribution displaying evidence of a strain toward normality, although obviously falling far short of that goal.

Samples of 40 should therefore continue the trend toward normality, or at least toward symmetry. Such a distribution is tabulated in Table 12.1.4b. Gone now is the long tail; instead, the beginnings of symmetry are clearly visible. Thus, sociologists who are often called upon to deal with skewed populations may draw considerable comfort from the important implications of this little experiment and may feel justified in

Table 12.1.4b

Distribution of 100 Sample Means, n = 40

Mean ('000)	*f*	Per Cent
100–199	0	0.0%
200–299	10	10.0
300–399	41	41.0
400–499	21	21.0
500–599	17	17.0
600–699	3	3.0
700–799	7	7.0
800–899	1	1.0
900–999	0	0.0
1,000 or more	0	0.0
	100	100.0%

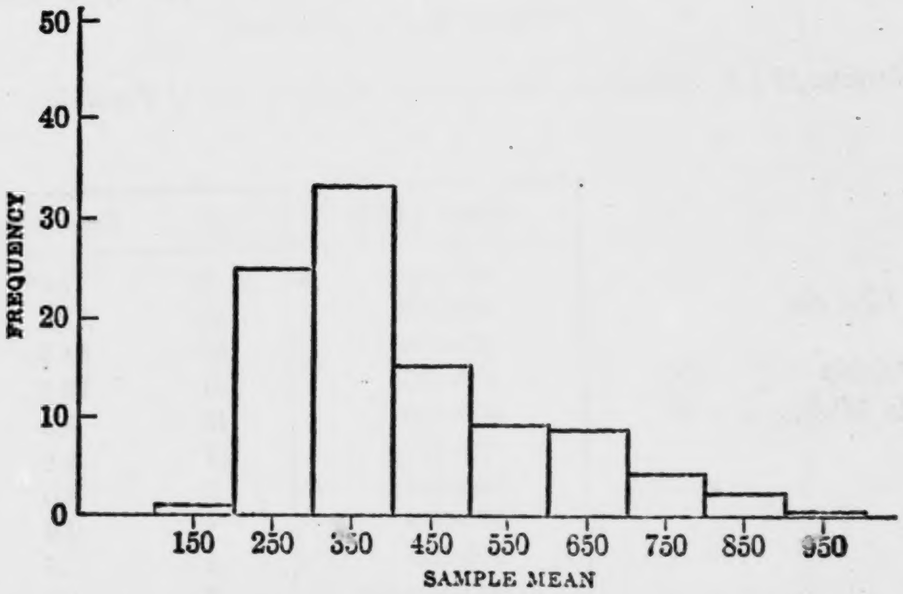


FIGURE 12.1.5a *Histogram, Distribution of 200 Sample Means, n = 20*

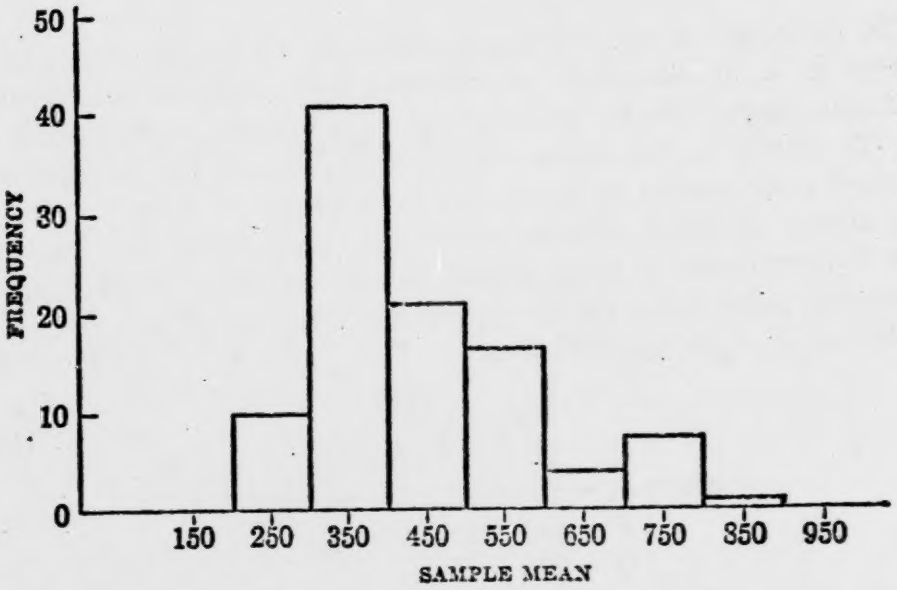


Figure 12.1.5b *Histogram, Distribution of 100 Sample Means, n = 40*

applying normal probability calculations to their data by taking proper precautions.

The convenience of large samples consists in the assurance that the sampling distribution of the mean is approximately normal, an assumption which may be in doubt when samples are small. Situations do, however, often arise when small samples are not only convenient, but even unavoidable. In such instances, modification of the foregoing procedures is required, although no new principles are involved. This adaptation of reliability techniques to small samples is therefore not considered in this text.

Confidence Interval for a Percentage (Large Sample). Like the mean, the sample percentage also raises the issue of reliability, since sample percentages are equally subject to sampling errors. Thus, a given sample percentage (e.g., 60 per cent favoring military training) may be off a trifle, or may deviate considerably from the true percentage. Similarly, 30 per cent of a random selection of married women may be gainfully employed, but the percentage in the sampled universe may conceivably be as low as 20 or as high as 40. Fifty-one per cent of the sample electorate may signify its intention of voting for the Republican nominee in the forthcoming election, but unless we can in some way assess the accuracy of that estimate we cannot forecast with any degree of confidence the outcome of the election. Hence, in interpreting a sample percentage, we must again estimate the average sampling error (standard error) in order to provide a confidence interval.

John H. Mueller and Karl F. Schuessler

The technique of constructing an interval estimate of a universe percentage is in its essentials no different from that of the mean: we (1) fix the desired degree of confidence; (2) estimate the standard error; and (3) attach to the observed sample percentage the multiple of the standard error needed to attain the agreed-upon confidence. The sampling theory on which this procedure rests is identical with that underlying the treatment of large sample means: namely, the distribution of all possible percentages (p) is approximately normal around the universe percentage, P, with standard error

$$\sigma_p = \sqrt{\frac{PQ}{n}} \tag{12.1.5}$$

where σ_p = standard error of percentage
P = universe percentage
Q = 100 − P

But since the parameters P and Q are of course unknown, we again follow the convention of substituting the sample values p and q in order to obtain an estimate of the standard error: *

$$s_p = \sqrt{\frac{pq}{n}} \tag{12.1.6}$$

Computing an Interval Estimate of a Percentage. Let us suppose that out of 900 randomly selected high school students, 60 per cent respond in favor of universal military training and 40 per cent are unfavorable. To form a confidence interval, we must first estimate the standard error:

$$\begin{aligned} s_p &= \sqrt{\frac{(60)(40)}{900}} \\ &= \sqrt{2.667} \\ &= 1.63 \end{aligned}$$

By adding and subtracting this quantity to and from the sample percentage, we obtain the 68 per cent confidence interval:

$$60\% \pm 1.63 \text{ percentage points}$$
$$58.37\% - 61.63\%$$

The odds are 2 to 1 that the true percentage lies within this interval. If we should desire greater confidence — say, 95 per cent — we would attach 1.96 standard errors:

$$\begin{gathered} 60 \pm 1.96(1.63) \\ 60 \pm 3.19 \\ 56.81\% - 63.19\% \end{gathered}$$

---

* As in the case of the mean, a better estimate of the standard error of p is provided by $\sqrt{\frac{pq}{n-1}}$. For a discussion of this point, see William G. Cochran, *Sampling Techniques*, John Wiley & Sons, Inc., New York, 1953, pp. 32–33.

Since intervals constructed in this manner contain the true percentage 95 times out of 100, we may declare that we are confident at the 95 per cent level that the true percentage is not less than 56.81, nor greater than 63.19. Since 5 per cent may be assumed to exclude the true percentage, we thus take a 5 per cent risk of being wrong.

The foregoing procedure is another instance of large-sample technique in that it applies to samples having 30 or more cases. But the assumption of a normal sampling distribution hinges on the ratio, $P:Q$, as well as on sample size. As the $P:Q$ balance departs more and more from 50:50, the sampling distribution becomes increasingly skewed.

As has already been pointed out in the study of the mean, in the case of percentages, the skew in the $P:Q$ balance communicates itself to the sampling distribution. Again, as with the mean, this skew can be circumvented only by a sharp increase in the size of n, so as to instill normality in the sampling distribution. Table 12.1.5 shows the sample sizes needed for varying sample percentages in order to permit the assumption of a normal sampling distribution.

Table 12.1.5

Minimum n for Selected p-Values for Normal Approximation

p	n
.5	30
.4	50
.3	80
.2	200
.1	600
.05	1,400

Adapted with permission from William G. Cochran, *Sampling Techniques*, p. 41. Copyright 1953. John Wiley & Sons, Inc.

The Finite Population Multiplier. In actual sociological investigations, we ordinarily encounter finite populations, and yet in the foregoing instances we have calculated the standard error on the assumption of sampling from an infinite universe. But this is not quite so unrealistic as might appear on its face. Finite populations are usually large enough to be considered as infinite, which is incidentally something of a practical convenience, since reliability procedures are somewhat less complex for infinite than for finite populations. For finite populations, the measurement of reliability must take into account both sample size (n) and the size of the universe (N), whereas for infinite populations, the computation of reliability need take into account only sample size, since the size of the universe is incalculable and hence cannot vary.

SECTION TWO

Testing a Statistical Hypothesis

Hypothesis-Testing and Estimation Compared. There are two general types of statistical inference: (1) *estimation,* which begins without any stated assumption about the value of the parameter and merely seeks to estimate descriptively what the parameter could be; and (2) *hypothesis-testing,* which begins with a hypothesis about the parameter and then uses the sample data to check the plausibility of that statement.

In the previous section, we were concerned with problems of estimation. We began, for example, with the observation of a sample mean, and from this we derived an interval estimate with a specified degree of confidence. We first drew our sample and then made our estimate of the parameter.

But in hypothesis-testing, we formulate our hypothesis about a parameter in advance of the collection of the sample data, which is then used to test that hypothesis. We may, for example, hypothesize that the average suicide rate of eastern cities does not differ from that of western cities. We begin with that supposition, and then we take an appropriate sample of eastern and western cities, compare the two means in the prescribed manner, and finally reach a decision whether, in the light of the sample difference, the hypothesis should be accepted or rejected.

Although an interval estimate is always derived from the previously compiled sample data, a statistical hypothesis involves quite another analytical process. This process starts with an antecedent conjecture about an unknown population value, presumably arrived at without benefit of the undrawn sample. Furthermore, a hypothesis is tested and then acted upon: it is either accepted or rejected. It requires a decision, which is made with a certain degree of confidence, and which is either correct or incorrect after it has been made. The emphasis thus shifts from mere estimation to *decision-making.*

The Null Hypothesis (H_0). It has become a convention in statistical testing to open the investigation with the null hypothesis, symbolized H_0. In its most current usage, this hypothesis holds that two or more given samples have come from statistically identical populations and that therefore any observed difference between such samples is merely a chance variation. The aforesaid hypothesis that East and West do not differ in their average suicide rates would therefore be a typical null hypothesis.

Essentially, the null hypothesis is set up to be nullified; however, every other type of hypothesis is also set up for that possibility. How, then, does the null hypothesis differ from any other hypothesis — from a hypothesis, for example, stating that juvenile delinquency is associated with status discontent, or that the incidence of suicide depends on the degree of social disorganization? These hypotheses we may call explanatory hypotheses, whereas the null hypothesis is an auxiliary device which tentatively denies the validity of the explanatory theory. Hence, if we reject the null hypothesis, we strengthen the credibility of our explanatory hypothesis. But the null hypothesis itself has no explanatory value.

Yet science is in search of relationships. Hence, in accordance with the above reasoning, the null hypothesis is usually launched with the expectation, indeed with the hope, that it will be nullified, as the derivation of the term implies. Nevertheless, it represents a possibility that must be disposed of before alternative hypotheses which imply assignable causes can be considered.

Returning to the comparison of eastern and western cities, we would set up such a comparison in the first place only in order to uncover some of the factors that determine the incidence of suicide. If we do discover a significant difference between the average rates of East and West and consequently reject the null hypothesis, we will have added a small increment to our knowledge of suicide — namely, the regional factor as a source of variation in the suicide rate. If, on the other hand, the two samples of data show no significant difference and we accept the null hypothesis, then we have not advanced our understanding, although supposedly we will not have retrogressed. In this latter case, we would have to contrive new kinds of comparisons in our search for the assignable causes of suicide.

Thus, it is only when the preliminary hypothesis of "no difference" has been cleared away that we gain some insight into the occurrence of suicide. Hence, the null hypothesis is inherently linked with a more constructive statistical hypothesis, sometimes called an alternative hypothesis, which becomes tenable to the extent that the null hypothesis has been discredited. Such an alternative hypothesis may specify an exact degree of difference between East and West, with a view to gauging the strength of the unknown variables which produce suicide in a disorganized society. Clearly, there is no point in pursuing the effects of a given control or experimental variable such as geographical region if we have found in favor of the null hypothesis. Thus, by analogy, the null hypothesis serves the purpose of a criminal trial: we set up the hypothesis of innocence, giving the evidence an opportunity to nullify it. Only if and when that presumption is rejected does the court give thought to alternative punishments corresponding to the degree of guilt.

In accordance with the foregoing principle, the null hypothesis has come to be predominantly identified with two types of research procedures: (1) the comparison of two or more populations on a given trait and (2) the correlation between two or more traits in a given population. In the first type, the null hypothesis posits no difference between population parameters; in the second type, it asserts a chance relation, or zero correlation, between the variables under study.

Principles of Testing the Null Hypothesis. Let us now suppose a random sample of 30 eastern cities and a comparable sample of 30 western cities whose mean suicide rates are 10.4 and 14.3, respectively (Table 12.2.1). We wish to know whether the observed difference of 3.9 could be indic-

Table 12.2.1

Suicide Rates, Eastern and Western Cities, $n = 30$

EAST	WEST
9	26
17	12
12	12
12	29
16	12
7	12
11	6
8	10
19	19
10	15
10	26
5	11
10	6
8	11
6	9
11	14
9	15
10	15
10	20
7	10
12	15
14	10
9	27
8	8
15	20
5	9
5	12
15	12
10	15
12	12
$\Sigma = 312$	$\Sigma = 430$
$\bar{X} = 10.4$	$\bar{X} = 14.3$

ative of an actual difference between the means of the sampled populations, or whether it could more plausibly be accepted as mere sampling variation. By rule of thumb, we lay down the null hypothesis that the population means of East and West are identical and subject this hypothesis to the statistical test. If we reject this hypothesis, we then proceed as though there were a real difference, whereas acceptance presumably implies that the two populations are alike and that the observed difference is a mere sampling error.

An important principle of our testing procedure is that *we can never prove the null hypothesis true.* The best possible evidence that we could ever obtain for the identity of the two population means would be an identity between the two sample means. But even if samples of East and West were to show identical averages, we still could not be positive that the null hypothesis was true. Such an identity between sample means could itself very well be the result of sampling errors, because population means could differ by a sizable amount and the sample difference still be zero. Similarly, the finding that a sample of married spouses are of the same average age would not prove that all married couples show equal averages. From sample observations, the conclusions we draw about the truth of the null hypothesis are necessarily of a probabilistic nature.

If we can never prove the null hypothesis true, may we prove it definitely false? Here the statistical evidence may be more convincing. The best possible evidence that the two populations differ would be a difference between samples. A small difference would not be very compelling; however, as this difference becomes larger and larger, the case for a population difference becomes stronger and stronger. If the difference between samples becomes so large that it is highly improbable that they stem from the same universe, then we may with practical assurance reject the null hypothesis. Even here, however, there is no infallible certainty, since the one case in a million — the freak event — may happen. Highly improbable events are regularly occurring, however startled we are when they befall us. But such an extreme case — the one in a million — is so improbable that most of us would discount it by regarding it as impossible, thereby rejecting the null hypothesis with great confidence. In other words, we are willing to reject the null hypothesis when the statistical probabilities of being wrong are small enough to suit our purposes.

These decisions — whether or not to reject the null hypothesis — are accordingly made with varying degrees of confidence. This confidence varies according to the probability that a difference at least as large as that observed could have been obtained by chance, assuming the truth of the null hypothesis. We must therefore now turn to the procedure by which such a probability is determined.

JURY SELECTION

The previous two readings provide law students with an introduction to statistical theory. The usefulness of coming to grips with such material is linked to two considerations: (1) there is a fundamental likeness between legal reasoning and statistical theory, and (2) law practice in the future will contain numerous demands that the practitioner be familiar with statistical procedures.

In discussing the similarity between legal reasoning and statistical theory Michael Finklestein argues that statistical decision theory is relevant to law because legal principles are the intuitive equivalents of ideas used in probability theory. While his illustration for this statement is based specifically on racial discrimination in jury selections, we feel that it has significance beyond his immediate example. "Probability theory," he notes, "has been found to apply to events commonly called random or chance." That is, the events are referred to as random because for all practical purposes there is no way of predetermining their outcome which is another way of saying that events have an equal chance of occurring. This, he concludes, is in theory similar to the constitutional requirement that jurors be chosen without regard to their race. This is not to say that juries must be entirely representative of the population with respect to the number of Negroes and whites selected. The use of literacy tests or using telephone listings may prevent Negroes from serving on juries; and, presumably, such devices do not violate the equal protection clause of the Constitution. But within the pool of eligibles, however restricted in size, the selection must be made without reference to race; and, as Finklestein points out, this means that the racial composition of jury venires is--or should be--the result of random selection and consequently an appropriate subject for probability theory analysis.

From Michael O. Finklestein, "The Application of Statistical Decision Theory to the Jury Discrimination Cases," Harvard Law Review, 80 (1966), pp. 338-376. Reprinted by permission.

Michael O. Finklestein*, "The Application
of Statistical Decision Theory to the
Jury Discrimination Cases," 80 Harvard
Law Review, pp 338-376 (1966).

Judges have often relied on intuitive notions of probability in deciding whether there has been discrimination in selection of juries. Mr. Finkelstein uses mathematical analysis to show the extreme improbability that nondiscriminatory selection occurred in situations in which courts have been unwilling to find a violation of the fourteenth amendment. His approach is not intended to supplant judicial weighing of values; on the contrary, it exposes important legal issues previously hidden by evidentiary difficulties.

THIS article will explore the applicability of the branch of mathematics known as statistical decision theory to cases involving claims of discrimination in the selection of jurors. It is my purpose to show that in these cases the Supreme Court has established legal principles which in many instances cannot rationally be applied without such mathematical techniques.

Over the past two hundred years the romance between mathematics and law has been for the most part a one-sided affair. The prospect of making their discipline effective in an area of social concern has persistently teased mathematicians, but lawyers, judges, and administrators have been slow to accept their advances. Elementary mathematical techniques occasionally have been used in the law where they obviously facilitated the achievement of legal objectives. Every judge deciding an antitrust merger case, for instance, has used simple mathematics — percentages — to illuminate share-of-market statistics. The reason is that the meaning of percentages is well understood and size comparisons are almost impossible without them. On a more sophisticated level, the apportionment of congressional representatives is now controlled by a mathematical technique known as the "method of equal proportions," a technique adopted because mathematicians and legal scholars were able to persuade Congress that the objectives of fair apportionment could best be accomplished in this way.[1] There is also an extensive literature on the subject of mathematical prediction of legal decisions by correlation analysis and other techniques.[2]

---

* Member of the New York Bar. A.B., Harvard, 1955, LL.B., 1958.

I am indebted to Professors Daniel Greenberg and Richard Friedberg, Columbia University, and to my wife Elinor, for their reading and correction of the manuscript.

It is a reflection of the empirical orientation of probability theory that this branch of mathematics was the first to be applied to a legal problem. In 1785, the French mathematician Marie Jean Antoine Nicolas Caritat Condorcet published his *Essai sur l'application de l'analyse à la probabilité des decisions rendues à la pluralité des voix.* Condorcet's *Essai* explored the probability of error in a decision made by a group of persons — the principal application of the theory being jury decisions. This subject captured the imagination of French mathematicians. Laplace discussed it in his *Traité des probabilités* published in 1812, and Poisson exhausted it in his *Récherches sur la probabilité des jugements en matière criminelle et en matière civile,* which appeared in 1837. No other legal problem has ever received such extensive mathematical treatment. In the introduction of his *Récherches* Poisson wrote hopefully that "among the applications of this calculus, one of the most important is that which concerns the probability of the correctness of judgments . . . ."[3]

The reception of this work would, perhaps, have surprised its author. Poisson's *Récherches* has become a cornerstone of modern probability theory, with an enormous range of scientific applications. Its predictive power has been demonstrated repeatedly. But the law itself has remained virtually immune to its influence. If there was any discernible impact on the French administration of justice, the fact has not been recorded, and from that day to this lawyers have made only slight use of probability and statistical theories although in certain respects these branches of mathematics are uniquely suited to legal problems.[4]

---

[1] 2 U.S.C. § 2(b) (1964). *See* Huntington, *The Apportionment of Representatives in Congress,* 30 Transactions of the Am. Math. Soc'y 85 (1928). For a mathematical approach to reapportionment see Weaver & Hess, *A Procedure for Nonpartisan Districting: Development of Computer Techniques,* 73 Yale L.J. 288 (1963).

[2] *See, e.g.,* Schubert, *Judicial Attitudes and Voting Behavior: The 1961 Term of the United States Supreme Court,* 28 Law & Contemp. Prob. 100 (1963). Unlike the mathematical analysis of legal issues which is proposed here, mathematical prediction frequently involves consideration of legally irrelevant facts (such as the type of court in which the case arises) which are found to have a statistical correlation with the result.

[3] S. Poisson at 1–2 (transl.).

[4] There has been some effort (much of it mistaken) to use statistical theory in connection with the so-called duplication problem. *See* 50 Minn. L. Rev. 745 (1966).

Michael O. Finklestein

Statistical decision theory is relevant to jury discrimination cases because legal principles in these cases are the intuitive equivalents of ideas used in probability theory. This does not imply that legal decisions can be the products of mathematical proofs; the differences between legal and mathematical reasoning preclude any such merging of the two sciences. It does mean, however, that in the more difficult jury discrimination cases an informed judicial decision cannot be made without mathematical analysis of the underlying data.

This article will present various methods of applying mathematical techniques to problems of jury discrimination, using as examples cases involving racial discrimination in the selection of veniremen and grand jurors. The same techniques can be applied to investigate discrimination on the basis of other factors, such as economic status or sex. A rather detailed mathematical treatment is presented here because it now seems clear from the fate of other efforts that mathematical techniques are unlikely to receive judicial acceptance until there is a much fuller understanding of the meaning of statistical methods and the nature of their relationship to the relevant legal issues.[5]

I.

For more than seventy-five years it has been the law that a conviction in a state court violates the equal protection clause of the fourteenth amendment if it is based on an indictment of a grand jury or a verdict of a petit jury from which Negroes were excluded because of their race.[6] Discrimination must of course be proved.[7] In the initial cases in which the principle emerged, problems of proof did not arise because the fact of discrimination was never put in issue. Negroes were excluded by statute; [8] the fact of exclusion was uncontested; [9] or the principle was tested solely on the pleadings.[10] In three early cases, discrimination was denied by the state, but the defendants failed to produce evidence to support their claims.[11]

---

[5] Attempts to apply statistical methods to these problems have been marred by cryptic explanations and have not met with judicial acceptance. Billingsley v. Clayton, 359 F.2d 13 (5th Cir. 1966), Brief for Appellants at 53–54 (probability of observed number of Negroes in the jury box said to be less than one in a billion in one division and less than one in 20 million in a second division); United States v. Hoffa, 196 F. Supp. 25 (S.D. Fla. 1961), Reply Memorandum for Defendants at 8 ("The statistical probabilities of random selection" made the absence of women "a virtual impossibility"). A far more elaborate effort was made in United States v. Van Allen, 208 F. Supp. 331 (S.D.N.Y. 1962). The defendants presented testimony by a statistician and a sociologist to the effect that the jury commissioners in the Southern District of New York discriminated against the poor. The court rejected the contention. Unfortunately the presentation was cryptic because the statistician did not reveal the basis for his conclusions.

It has been suggested that probability theory might be used to predict or explain the outcome of jury discrimination cases. Ulmer, *Supreme Court Behavior in Racial Exclusion Cases: 1935–1960*, 56 Am. Pol. Sci. Rev. 325 (1962). The method presented here is intended not to reduce past holdings to a mathematical standard but to provide courts with a new method of analyzing evidence.

The current type of case first appeared in 1934 with the Supreme Court decision in *Norris v. Alabama*.[12] There the Court faced for the first time a disputed issue of discrimination in the selection of grand and petit jurors and a record containing evidence purporting to bear on the issue. The evidence on the motion to quash the indictment was that although Negroes comprised 7½% of the total adult male population of Jackson County, where the petitioners were indicted, no Negroes in the recollection of witnesses had ever served on any grand or petit jury. This single fact, the Court held, made out a "*prima facie* case of the denial of the equal protection which the Constitution guarantees." [13] The prima facie case was "supplemented," the Court noted, by evidence of the presence of at least thirty qualified Negroes in the county and by the testimony of the jury commissioners that Negroes had not been included on the jury roll because their names were "never discussed."

The evidence on the motion to quash the trial venire was that although Negroes comprised about 18% of the adult male population in Morgan County, where the petitioners were tried,[14] and although many were qualified, no Negro within the memory of witnesses had ever served on a jury or been called for service. The county sheriff was unable to identify any names of Negroes on the jury roll. The Court held the general denials of discrimination by a member of the jury board insufficient "to rebut the strong *prima facie* case which the defendant had made."[15]

---

[6] The seminal cases are Strauder v. West Virginia, 100 U.S. 303 (1880), and Neal v. Delaware, 103 U.S. 370 (1880). Legislation designed to enforce these rulings has recently been rejected by the Senate. 112 CONG. REC. 22,091–117 (daily ed. Sept. 19, 1966).

[7] Tarrance v. Florida, 188 U.S. 519 (1903).

[8] Bush v. Kentucky, 107 U.S. 110 (1883).

[9] Neal v. Delaware, 103 U.S. 370 (1881).

[10] Carter v. Texas, 177 U.S. 442 (1900).

[11] Martin v. Texas, 200 U.S. 316 (1906); Brownfield v. South Carolina, 189 U.S. 426 (1903); Smith v. Mississippi, 162 U.S. 592 (1896).

[12] 294 U.S. 587 (1935).

[13] *Id.* at 591.

[14] The trial was transferred to Morgan County because of asserted prejudice against the defendants in Jackson County.

The issue in *Norris* was whether the particular grand and petit juries which indicted and convicted the petitioners were selected without discrimination. No testimony was introduced pertaining specifically to the selection of either jury; instead, discrimination was proved from the racial statistics of jury selection over a number of years. Statistical evidence of this type can be relevant in different ways to the existence of discrimination in the choice of a particular venire. The absence of Negroes from other venires selected from the same jury roll may be evidence that there are no Negroes on that roll. The practice of the jury commissioners in excluding Negroes from other venires may be evidence of their intent to exclude Negroes from the venire in question.[16] The unvarying failure of different groups of commissioners to select any Negroes may be evidence of a governing custom of discrimination. Whatever the relevant inference, the deduction of discrimination rests on the idea that the absence of Negroes from the jury roll or from jury venires is evidence of racial exclusion somewhere in the selection process when Negroes constitute a substantial segment of the population.

The facts in *Norris* made it easy for the Court to accept this idea. The absence of Negroes from juries, jury venires, and the jury roll, and the direct testimony of the commissioners that Negroes were never considered, combined to present a clear case of discrimination. The facts in other cases in which the Court has granted relief have been equally extreme.[17] But since, on the Court's reasoning, the Constitution prohibits *any* discrimination in selection, it was inevitable that the ideas used by the Court in *Norris* would be tested in cases where the evidence was less blatant. Challenges were made in cases where Negroes had appeared on juries, although in proportions smaller than their proportion in the population. In such underrepresentation — as opposed to exclusion — cases, the Supreme Court has yet to hold that the disparity was sufficient to show discrimination.[18]

---

[15] 294 U.S. at 598.

[16] In Brown v. Allen, 344 U.S. 443, 479 (1953), the Court noted that "past practice is evidence of past attitude of mind." It found such evidence not decisive because there had been a purge and refilling of the jury box before the jury in question was selected.

[17] *E.g.*, Arnold v. North Carolina, 376 U.S. 773 (1964) (one Negro served on grand jury in 24 years); Hill v. Texas, 316 U.S. 400 (1942) (no Negroes served on grand jury for at least 16 years).

In all, five to eight underrepresentation cases have come before the Court.[19] There is promise of many more. In addition to a larger flow of cases involving criminal defendants, an increasing number of class actions has been brought on behalf of Negroes to secure the right to serve on juries.[20] The plaintiffs in these cases seek not to set aside a conviction but to restructure the selection system. Inasmuch as the claims in these cases involve many juries, the proof by necessity is statistical. The statistics usually show not exclusion but underrepresentation, thus testing the principles of the exclusion cases in an important institutional setting.[21]

Dallas County, Texas, was the source of the first two underrepresentation cases. The stage was set in 1942 when the Supreme Court, in *Hill v. Texas*,[22] reversed the murder conviction of a Negro because Negroes had been totally excluded from grand juries, including the grand jury which indicted him. Following this decision, a Negro appeared on each of the next two grand jury lists, and one served on the grand jury which indicted Robert Akins, a Negro, for the rape of a white woman. Akins appealed his conviction to the Supreme Court, claiming that the number of Negroes serving on the grand jury which indicted him had been limited by design to one. In *Akins v. Texas*[23] the Court rejected this claim. It computed that "on the strictly mathematical basis of population, a grand jury of twelve would have 1.8552 Negro members on the average," apparently deriving this figure by multiplying the Negro percentage of the population by the number of jurors. Since there had been one Negro member on each jury list, and since one served on the jury that indicted Akins, the Court concluded that "we cannot say that the omission from each of the two lists of all but one of the members of a race which composed some fifteen per cent of the population alone proved racial discrimination."[24]

---

[18] *But cf.* Anderson v. Alabama, 366 U.S. 208 (1961), *rev'g per curiam* 270 Ala. 575, 120 So. 2d 397 (1960), where both underrepresentation on venires and exclusion from grand juries were alleged.

[19] Akins v. Texas, 325 U.S. 398 (1945); Cassell v. Texas, 339 U.S. 282 (1950); Brown v. Allen and Speller v. Allen, 344 U.S. 443 (1952); Swain v. Alabama, 380 U.S. 202 (1965). The number is eight if Smith v. Texas, 311 U.S. 128 (1940), Barksdale v. Louisiana, 247 La. 198, 170 So. 2d 374 (1964), *cert. denied*, 382 U.S. 921 (1965), and Anderson v. Alabama, 366 U.S. 208 (1961), are included. Coleman v. Alabama, 377 U.S. 129 (1964), has been remanded for hearing on the merits of the underrepresentation issue. Sims v. Georgia, 221 Ga. 190, 144 S.E.2d 103 (1965), *cert. granted*, 384 U.S. 998 (1966), is pending.

[20] *E.g.*, White v. Crook, 251 F. Supp. 401 (M.D. Ala. 1966) (jury box required to be emptied and refilled).

[21] *See, e.g.*, Mitchell v. Johnson, 250 F. Supp. 117 (M.D. Ala. 1966), where the number of white persons on the jury roll represented 55% of the eligible white population and the number of Negroes represented only 9% of the Negro population. The court held that this evidence demonstrates "wide disproportions" and that this, "without more, requires an inference of systematic exclusion on racial grounds . . . ."

[22] 316 U.S. 400 (1942).

[23] 325 U.S. 398 (1945).

During the five and one-half years after the *Hill* decision, there were twenty-one grand juries in Dallas County. It is consistent with the implied invitation in *Akins* that one Negro each appeared on seventeen grand juries and none on the remaining four. These statistics were scrutinized by the Supreme Court in *Cassell v. Texas*,[25] the third of the Dallas County cases. Cassell, a Negro convicted of murder, charged discrimination and relied on the above statistics and on the testimony of the jury commissioners to prove it. Since there had been only seventeen Negroes of the 252 members on the twenty-one grand juries, only 6.7% of the jurors were Negroes. The Court observed, however, that the discrepancy between this figure and the 15.5% Negro population was explicable, because Negroes constituted only 6.5% of the county poll-tax payers and payment of the poll tax was a prerequisite for jury service. Since 6.7% was the ratio of Negroes actually sitting on the juries, the Court concluded that "without more it cannot be said that Negroes had been left off the grand jury panels to such a degree as to establish a prima facie case of discrimination."[26]

The *distribution* of Negroes presented, in the Court's view, a different question. If it was true that the jury commissioners had limited their number to one on each jury, this action would have been unconstitutional since "jurymen should be selected as individuals . . . and not as members of a race."[27] But Mr.

---

[24] *Id.* at 405-06. Apart from the superficiality of the mathematical argument, the decision might not have caused confusion, but the Court added a *caveat* to its opinion which is wholly inexplicable in the light of former rulings that race may not be considered in the selection process: "This conclusion makes it unnecessary to decide whether a purposeful limitation of jurors by race to the approximate proportion that the eligible jurymen of the race so limited bears to the total eligibles is invalid under the Fourteenth Amendment." *Id.* at 407.

[25] 339 U.S. 282 (1950).

[26] *Id.* at 285-86. This is a classic example of the failure of intuition to comprehend the significance of data.

[27] 339 U.S. at 286.

Justice Reed's opinion for the Court, concurred in by three other Justices, avoided resting decision on the ground that the appearance of a single Negro on each of seventeen juries was evidence of limitation; instead, it reversed because the commissioners testified that they took only people they knew and admitted they did not know any Negroes. "When the Commissioners were appointed as judicial administrative officials, it was their duty to familiarize themselves fairly with the qualifications of the eligible jurors of the county without regard to race and color. They did not do so here, and the result has been racial discrimination."[28]

Although the decision in *Cassell* was based on other grounds, the case was the first in which the Court accepted an explanation for a disproportionately small Negro representation on juries. The issue reappeared a few years later in two cases reported together from North Carolina, *Brown v. Allen* and *Speller v. Allen*.[29] The Court then proved willing to accept such explanations. Brown and Speller, Negroes, were convicted in the North Carolina courts of rape and were sentenced to death. Their challenges to the grand and petit juries were based solely on statistical evidence. In Brown's case this showed that, although Negroes comprised 33.5% of the adult population of the county, in the year of his indictment and the previous year only 7% to 10% of those chosen for grand jury service were Negroes; in the year he was indicted the percentage of Negroes on petit jury panels ranged from 9% to 17%. The Court stated that "variations in proportions of Negroes and whites on jury lists from racial proportions in the population have not been considered violative of the Constitution where they are explained and not long continued."[30] It concluded that the disparity had been explained. The names of prospective jurors were drawn from poll and property tax lists, and Negroes appeared to constitute a much smaller percentage of such lists than of the general population — although how much smaller a percentage did not appear from the evidence. The decision below was affirmed.

In Speller's case, jurors were selected from the tax lists. Thirty-eight percent of the taxpayers were Negroes, but only 7% of the persons selected by the clerk for the jury box were Negroes. The Court held, however, that the discrepancy could be explained by the testimony of the clerk that he took those with "the most property." It concluded that "evidence of discrimination based solely on race in the selection actually made is lacking." [31]

---

[28] *Id.* at 289.

[29] 344 U.S. 443 (1953).

[30] *Id.* at 471.

The apparent reluctance of the courts to consider underrepresentation as evidence of discrimination has been confirmed by two recent cases, which will be considered in some detail in Part III. In the first case, *Swain v. Alabama*,[32] Swain, a Negro, was convicted of rape in Talladega County, Alabama, and sentenced to death. The Supreme Court summarized the evidence on the panels as follows: "While Negro males over 21 constitute 26% of all males in the county in this age group, only 10 to 15% of the grand and petit jury panels drawn from the jury box since 1953 have been Negroes, there having been only one case in which the percentage was as high as 23%." [33] In Swain's case there were four or five Negroes on the panel of thirty-three from which his grand jury was selected and eight Negroes on the panel of one hundred from which his petit jury was selected. The Court held that these statistics did not constitute prima facie evidence of discrimination: "We cannot say that purposeful discrimination based on race alone is satisfactorily proved by showing that an identifiable group in a community is underrepresented by as much as 10%. . . . The overall percentage disparity has been small, and reflects no studied attempt to include or exclude a specified number of Negroes." [34] As we shall demonstrate, this result is inconsistent with the announced principles governing these cases.

Swain also attacked the petit jury which had convicted him, claiming that the prosecution had used its peremptory strikes to exclude Negroes. The evidence showed that the six Negroes available for service in Swain's case had been struck by the prosecution. The Court, three Justices dissenting, rejected the claim. It held that in any given case it is presumed that the prosecution has used its peremptory challenges to secure an impartial jury and that this presumption is not overcome by allegations that "in the case at hand all Negroes were removed . . . because they were Negroes." [35] In the Court's view the case would have been different, and the fourteenth amendment claim would have taken on added significance, if defendant had shown that the prosecutor "in case after case, whatever the circumstances, whatever the crime and whoever the defendant or the victim may be, is responsible for the removal of Negroes who have been selected as qualified jurors by the jury commissioners and who have survived challenges for cause . . . ."[36] But although Negroes had never served on petit juries in Talladega County, Swain had failed to prove his case because he had not shown that the prosecution had been responsible in each case for striking Negroes.

---

[31] *Id.* at 481.

[32] 380 U.S. 202 (1965).

[33] *Id.* at 205.

[34] *Id.* at 208–09. The deceptiveness of this statement was called to the Court's attention. Brief of the American Civil Liberties Union as Amicus Curiae in Support of Petition for Rehearing at 10–11. If 90% of the population and 80% of the average panel is Negro, the "10% disparity" may be insignificant. This is not so if 10% of the population is Negro but *no* panelists are Negro.

[35] 380 U.S. at 222.

The Court has thus drawn the line between a peremptory strike based on a belief that a Negro juryman would be biased in a particular case and the systematic use of peremptory strikes to keep Negroes off juries. The first is permitted, the second unconstitutional.[37]

In the other recent case, *State v. Barksdale*,[38] a Negro was convicted of rape and sentenced to death. The evidence showed that Negroes in Orleans Parish composed about 33% of the adult male population but only 10% to 16% of the grand jury venires and 11% to 19% of the petit jury venires. From May 1958, when the Supreme Court reversed a murder conviction in the parish on the ground of exclusion of Negroes from the grand jury,[39] until September 1962, when Barksdale was indicted, nine grand juries had been selected. Eight of these contained two Negroes each, and one Negro served on the ninth.[40] On this evidence the Supreme Court of Louisiana affirmed, holding that the smaller proportion of Negroes serving on grand juries was adequately explained by their deficiencies in education (which resulted in the qualification of fewer Negroes) and income (which led more Negroes to request hardship excuses).[41] The Supreme Court denied certiorari. The denial, of course, does not mean that the Court approved the state court's holding. I shall try to show that, when put to a mathematical test, the facts in *Barksdale* point to a contrary decision.

The opinions of the Court in underrepresentation cases leave little doubt that disparity between the proportion of Negroes on venires and in the population generally is evidence of discriminatory selection if it is large, continuing, and unexplained. But it is also clear that a defendant is not entitled to a venire or a jury on which members of his race are proportionately represented, and efforts to create proportional representation by systematically including Negroes have been held unconstitutional.[42] Consequently, not every disparity in proportions is constitutionally fatal. "The mere fact of inequality in the number selected does not in itself show discrimination." [43]

---

[36] *Id.* at 223.

[37] *See* Note, *Fair Jury Selection Procedures*, 75 YALE L.J. 322 (1965).

[38] 247 La. 198, 170 So. 2d 374 (1964), *cert. denied*, 382 U.S. 921 (1965).

[39] Eubanks v. Louisiana, 356 U.S. 584 (1958).

[40] 247 La. at 223, 170 So. 2d at 383.

[41] *Id.* at 223–24, 170 So. 2d at 384.

The refusal of the Court to grant relief in *Swain* has thrown into sharp relief the problem of drawing a line between significant and insignificant disparities. Counsel in *Swain* asserted flatly that "the State must surely offer some explanation of why the proportion of Negroes on grand and petit jury venires averages at most one-half the proportion of eligible Negroes in the population." [44] But neither Court nor counsel offered a rational way of determining which disparities were large enough to be considered evidence of design. As a result, counsel could do little more than assert, without reasoned support, that the statistics required a finding of discrimination. And when the Court reached a contrary conclusion in *Swain*, the basis for its decision remained equally Delphic.

In the exclusion cases the Court has consistently refused to countenance the explanation that Negroes were omitted from juries because they were unqualified. The explanation arose again in the underrepresentation cases, this time in more compelling form, since all that had to be asserted was that some but not all Negroes were unqualified. In *Cassell*, *Brown*, and *Speller*, the Court accepted such assertions, refusing to find discrimination on the basis of disparities in percentages. In *Cassell* the evidence showed that the average proportion of Negroes on grand juries was at least approximately equal to their proportion on the tax rolls; but in *Brown* and *Speller* the evidence showed only that the number of eligible Negroes was to some undetermined extent smaller than their number in the adult male population. On what theory the Court found nonquantitative explanations sufficient does not appear.

The failure of the Court to articulate a rationale for the underrepresentation cases is a consequence of the problem they present.

---

[42] *E.g.*, Collins v. Walker, 329 F.2d 100, *aff'd on rehearing*, 335 F.2d 417 (5th Cir.), *cert denied*, 379 U.S. 901 (1964).

[43] Akins v. Texas, 325 U.S. 398, 403 (1945).

[44] Brief for Petitioner at 19.

In the exclusion cases the disparity is gross enough for the intuitive conclusion that jurors were chosen on the basis of race. But the notion that *some* Negroes should have been chosen is too imprecise to offer guidance once some Negroes are in fact chosen. Since the judges have recognized that not every variation from proportionate numbers is evidence of design, they are confronted with the harder problem of drawing a line between significant and insignificant variations and the related problem of determining the effect to be given to "explanations" for significant variations. The resolution of these problems involves important questions of legal policy which the Court has never reached because the legal issues have remained intertwined with statistical ones. We need not more tender-minded Justices or more appealing verbal formulations but an analysis of the data which will expose the true legal issues.

II

Probability theory has been found to apply to events commonly called "random" or "chance." The result of a coin toss is the classic example. Theoretically, at least, the person tossing the coin might determine the outcome in advance by calculating the forces acting on the coin. But, since minute changes in some of these forces can produce a different outcome, there is for all practical purposes no way of predetermining the outcome. Consequently, the result is treated as independent of the volition of the actor. It is this practical independence of result from cause that characterizes phenomena suited to the application of probability theory.[43]

It is obvious that the racial outcome of jury selection is — or rather should be — independent of the method of selection. This is true not because the selecting agent is unable to control the result, as in the coin-tossing case, but because he is not legally permitted to do so. But whether the outcome is independent because of the physical impossibility of connecting it with its causes or because of the unconstitutionality of making such a connection, the result is the same: the racial outcome is random.

This is not to say that the method of selection for the jury roll must be neutral with respect to the number of whites and Negroes considered. The use of various qualification tests, such as literacy, or the operation of the selection machinery, for instance the use of telephone listings, may result in the qualification or consideration of a disproportionately small number of Negroes. But within the pool of eligibles, however restricted in size, the selections must still be made without reference to race, and this means that the racial population of jury venires is — or should be — the result of random selection and consequently an appropriate subject for probability theory analysis.

---

[43] The possibility of perfect prediction is doubtful even in theory because of the teaching of quantum theory that certain events involving fundamental particles have no determinable cause.

Basing itself on probability theory, statistical decision theory provides us with a variety of tests for the hypothesis that the number of Negroes chosen for venires is consistent with random selection. Like all statistical tests, these are not infallible, but their rates of error can be made extremely small and are mathematically determinable. There is, as we shall show, judicial precedent for deciding cases with a rate of error similar to the rate of error present in the tests we propose to use.

The Supreme Court has used an approach in principle identical to this when holding that the persistent absence of Negroes from juries is evidence of discrimination. Although the basis for decision has usually not been articulated, the Court has indicated on a few occasions that its holdings rest on the improbability of repeated random selection of all white venires in counties with substantial Negro populations. Thus, in *Smith v. Texas*, the Court said: "Chance and accident alone could hardly have brought about the listing for grand jury service of so few negroes from among the thousands shown by the undisputed evidence to possess the legal qualifications for jury service."[46]

Although stated as a single rule, the Court's theory in these cases involves two propositions. The first is that, if the number of Negroes chosen is an improbable result of random selection, that fact is prima facie evidence of discrimination. This approach justifies our use of probability and statistical analysis in these cases. As we shall see, improbability may be demonstrated in a number of different ways. The Court has explicitly sanctioned one of these ways in its second proposition: a large and unexplained disparity between the proportion of Negroes on venires and in the population generally is unlikely to result from random selection. This approach justifies the mathematical model which we shall use to determine the probabilities in these cases.

The mathematical test which can be used to determine the existence of discrimination is based on the probability that, as veniremen are selected, any given venireman will be a Negro. This probability is usually derived from the proportion of adult Negro males to the total adult male population. In each jury case, however, it has been urged that these population figures are misleading and that the absence of Negroes from venires or juries reflects a lack of literacy, of integrity, or of other qualifications required of jurors. Statistical methods can be fruitfully used here to determine the percentages of Negroes and whites it would be necessary to assume disqualified in order to account for the observed numbers of Negroes on venires. Whether these percentages are consistent with the education, status, and general qualifications of the Negro and white populations is a question which the Court can approach with some confidence since it resembles other factual issues underlying constitutional claims.

---

[46] 311 U.S. 128, 131 (1940); *accord*, Eubanks v. Louisiana, 346 U.S. 584, 587 (1958); Hill v. Texas, 316 U.S. 400, 404 (1942).

The Court has taken a similar but nonquantitative approach to the qualifications issue in the cases we have discussed. In the exclusion cases, the contention that the absence of Negroes from juries could be explained by their lack of qualifications was rejected because it involved the assumption (so the Court believed) that virtually the entire Negro population was unqualified.[47] "[A] race [cannot] be proscribed as incompetent for service."[48] Similar reasoning appears in the underrepresentation cases. In *Speller v. Allen*,[49] for example, the Court held that a discrepancy of thirty-one points between the percentage of Negroes on venires and the percentage on tax lists from which the venires were drawn could not be explained by the clerk's disqualification of Negroes for poor moral character. "It would not be assumed," the Court held, "that in Vance County there is not a much larger percentage of Negroes with qualifications of jurymen."[50] The same method of reasoning, using mathematical techniques, will be adopted here.

In certain situations it is impossible to determine whether the number of Negroes appearing on a series of venires or juries is consistent or inconsistent with the qualifications of the Negro population. This may occur when there is inadequate evidence concerning qualifications or when persons selected for service may be excused on grounds other than lack of qualifications. Those called for grand jury service, for instance, may be excused on grounds of economic hardship. Another example is the appearance of Negroes on petit juries (as distinguished from jury panels). Although in *Swain* the Supreme Court stated that the prosecution may not use peremptory strikes to exclude Negroes systematically, there remains a broad area of discretion in the use of such strikes or challenges which makes it virtually impossible to determine from population statistics, however carefully refined, the probability that a Negro will appear on a petit jury.

---

[47] Patton v. Mississippi, 332 U.S. 463, 468 (1947).

[48] Brown v. Allen, 344 U.S. 443, 471 (1953).

[49] 344 U.S. 443 (1953).

[50] *Id.* at 481.

While in such cases it is not possible to use a statistical analysis based on the failure to select a greater number of Negroes, it is possible, at least in the case of venires and grand juries, to determine whether the number of Negroes selected, which I shall call the *distribution* of Negroes, is consistent with randomness. To take a crude example, the appearance of a series of grand juries alternately 100% and 0% Negro would excite our suspicions as to the selection process whatever the probability of selecting a Negro or white juror. Statistical decision theory provides a measure of the extent to which a given distribution differs from the theoretical or expected random distribution and a test of the significance of such difference. The test will indicate, within mathematically determined limits of accuracy, whether the nature of the observed distribution is so at variance with the expected distribution that the hypothesis of random selection ought to be rejected.

Although the Court has never applied this line of reasoning, the test of distribution operates on the same principles as the test based on a comparison of percentages which it has used. Perhaps the closest thing we have to judicial recognition that a skewed distribution may be evidence of discrimination is the statement of the three concurring Justices in *Cassell v. Texas*:[51]

> The number of Negroes both qualified and available for jury service in Dallas County precluded such uniform presence of never more than one Negro on any other basis of good faith than that the commissioners were guided by the belief that one Negro on the grand jury satisfied the prohibition against discrimination . . . .

In concluding that probability theory can usefully be applied to these problems, it has been assumed that the intuitive idea of probability used by the Court is conceptually similar to the mathematical definition used by the statisticians. Although the premises of the legal argument remain obscure, the assumption is a reasonable one, since the intuitive content of the mathematical definition is the common idea of probability. Thus the mathematical notion of an event with a probability of, say, 0.1 corresponds to the practical notion of an event that occurs on the average of once in every ten trials.[52]

---

[51] 339 U.S. 282, 294 (1950) (Frankfurter, J., concurring).

III

We now apply the methods outlined above to *Swain v. Alabama* and *Barksdale v. Louisiana.* The first case involves the selection of Negroes for venires; the second, their appearance on grand juries.

A. *An Analysis of the Selection of Veniremen*

In *Swain*, it will be recalled, there was a claim of discrimination in the selection of the grand and petit jury venires and in the selection of petit jurors. We begin our examination with a general mathematical analysis of venire selection. The application of these methods to the facts in *Swain* will be confined to the grand jury venires, since the record with respect to the petit jury venires is not sufficiently clear to permit the use of mathematical methods.

The process of selecting veniremen can be treated as what is known in probability theory as a series of Bernoulli trials.[53] Repeated trials are called Bernoulli trials if (*1*) there are only two possible outcomes for each trial, (*2*) the trials are independent, and (*3*) the probabilities for each outcome remain constant throughout the trials. The repeated tossing of a coin is a series of Bernoulli trials because there are, theoretically, only two outcomes for each toss, a head or a tail, the tosses are independent in the sense that the outcome of one does not affect the outcome of another, and the probability of tossing a head or a tail remains constant throughout any series of tosses. Similarily, when veniremen are selected, there are only two racial results possible, white and nonwhite (we equate the latter with Negro), the result of each selection is independent of other selections, and it is reasonable to assume that the probabilties of selecting a Negro or a white remain constant throughout the selection process.

Strictly speaking, selecting veniremen is not a series of

---

[52] For a nontechnical discussion see Nagel, *The Meaning of Probability*, in 2 THE WORLD OF MATHEMATICS 1398 (J. Newman ed. 1956).

[53] Named after Jacob Bernoulli (1654-1705), whose famous treatise on probability, *Ars Conjectandi*, appeared posthumously in 1713.

Bernoulli trials because each person selected is eliminated at least for a certain time from the pool of potential jurors. Theoretically, therefore, the probability of selecting a white or a Negro on any given trial will be dependent on the history of selection to the extent that the racial proportions of the available population are changed by the elimination of those already chosen. Although it is possible to take this into account in computing probabilities, it is not necessary to do so. The number of veniremen selected will in most cases change the racial proportions of the available population to an insignificant degree. Moreover, in these cases whites are usually selected with a frequency which equals or exceeds their proportion of the population. The assumption that the proportion of white and Negro eligibles remains constant is conservative since the proportion of white eligibles actually declines during the selection process. Assuming that the proportion remains constant makes an improbably large percentage of white veniremen seem slightly more in accord with random selection than is really the case. For these reasons, we adopt the simplification that the selection process consists of independent trials with constant probabilities applicable to each trial.

The derivation of the probabilities associated with Bernoulli trials rests on two basic propositions. The first is the sum rule, which states that the probability of the occurrence of any one of a number of mutually exclusive events is equal to the sum of the probabilities of those events.[54] The second is the product rule, which states that the probability of the joint occurrence of a number of independent events is equal to the product of the individual probabilities.[55] Discussion of these elementary propositions may be found in textbooks on probability or statistics.[56]

Applying the product rule, if P is the probability of selecting a single Negro venireman, the probability of selecting two Negro veniremen is $P \times P$, or P^2. More generally, the probability of selecting v Negro veniremen, where v is any number, is P^v. Similarly, if Q is the probability of selecting a single white venireman,[57] then the probability of selecting w white veniremen is Q^w. If $Q = 3/4$, as it may if whites comprise approximately three-quarters of the eligible population, and if there are thirty veniremen on a panel, the chance of picking a single panel without any Negroes is, by the product rule,[58] $Q(w = 30) = (3/4)^{30} = 0.0002$. This means that on the average only two venires in ten thousand would be all white. If every venire were all white or all Negro our analysis could stop here. But, since in *Swain* and most other cases both whites and Negroes were in fact chosen, some further analysis is necessary to derive the relevant probabilities.

---

[54] Events are mutually exclusive if the occurrence of one precludes the occurrence of the other. For example, if we consider the selection of two veniremen and define the first event as the selection of a Negro on the first but not on the second trial, and define the second event as the selection of a Negro on the second but not on the first trial, then the two events are mutually exclusive since they cannot both occur in the same two drawings.

[55] Events are independent if the occurrence of one has no effect upon the occurrence of the other. Thus, whether or not a Negro is selected as a venireman in one drawing should have no effect upon the selection of a Negro as a venireman in the next drawing. The product rule is frequently used as a definition of independence.

[56] *See, e.g.*, P. HOEL, INTRODUCTION TO MATHEMATICAL STATISTICS 8–12 (3d ed. 1962).

If P is the probability of selecting a single Negro and Q the probability of selecting a single white, PQ is, by the product rule, the probability of selecting a Negro and a white in that order. The order of selection, however, is unimportant. The condition of selecting a Negro and a white will also be satisfied by first selecting a white and then selecting a Negro. The total probability of selecting a Negro and a white without regard to order is thus $PQ + QP = 2PQ$. In general, the probability of selecting a particular mixture of Negroes and whites is the probability of selecting that mixture in a particular order times the number of different orders in which the mixture may be selected.[59]

If n is the number of men on a venire, the probability of selecting v Negroes is P^v and the probability of selecting $n - v$ whites is $Q^{(n-v)}$. Since these events are independent, their product, $P^vQ^{(n-v)}$, is the probability of selecting v Negroes and $n - v$ whites in a particular order. The second element is the number of ways this mixture may be selected. The number of distinguishable orderings of n objects of which v are of one class and $n - v$ of another class is: [60]

---

[57] Since we assume that "Negro" is equivalent to "nonwhite," either a white or a Negro must be selected. Consequently $P + Q = 1$.

[58] The symbol P or Q followed by $(v = a)$ or similar expressions is used to mean the probability of selecting a venire with a Negroes (or whites). Variations of this expression should be understood accordingly.

[59] Thus if $P = 1/4$ and $Q = 3/4$, then $PQ = 3/16$ and $2PQ = 6/16$.

[60] The derivation of this formula is not difficult. If there are n distinguishable objects, the number of possible orderings is n factorial, which is written as $n! = (n)(n-1)(n-2)\ .\ .\ .\ (1)$. The reader can satisfy himself that this is so by observing that there is one ordering of one object and that the addition of the nth object always multiplies the number of orderings by n, since for each previous ordering there are now n orderings corresponding to the n possible positions of the nth object when it is inserted among the $n - 1$ objects already present. Hence with the addition of the second object there are 2×1 orderings. The formula $n!$

$$\binom{n}{v} = \frac{n!}{v!\,(n-v)!}$$

These two elements (the probability for each ordering and the number of distinguishable orderings) multiplied together yield the well-known "binomial distribution," which expresses the probability $P(v)$ of selecting v Negroes out of n veniremen:

$$P(v) = \binom{n}{v} P^{v}Q^{n-v}$$

The challenge to the selection system we are discussing here must be based, not on the probability of selecting a particular number of Negroes (which is what the above formula gives us), but on the probability of selecting not more than i Negroes. By the sum rule, the probability of selecting not more than i Negroes is the sum of the probabilities of selecting 0, 1, 2, . . . i Negroes. This sum is expressed symbolically as follows:

$$P(v \leqq i) = \sum_{v=0}^{v=i} \binom{n}{v} P^{v} Q^{n-v}$$

With this preparation, we are ready to consider the grand jury venires in *Swain*. The record shows that for about ten years Negroes had never accounted for more than 15% of the grand jury venires, with the single exception of a 23% Negro venire chosen in the summer of 1955. These venires consisted of approximately 30 persons each, and three to four were chosen every year.[61] Census figures showed that adult Negro males constituted approximately 26% of the adult male population. For purposes of discussion we assume that between 1955 and 1962 thirty venires were chosen consecutively with five or fewer Negroes each, and that between 1953 and 1955 five additional venires were selected

---

follows immediately by induction. 0! is defined as equal to 1. If v of the n objects are the same, then the number of orderings must be divided by $v!$ since — given any ordering — one can rearrange the v indistinguishable objects in $v!$ ways without obtaining a distinguishable ordering. In other words, one can divide the $n!$ total orderings into $n!/v!$ groups of $v!$ orderings each, and the $v!$ orderings in any group are indistinguishable, so that there are only $n!/v!$ distinguishable orderings. Consequently, if v objects are of one type and $n - v$ objects of another type, the number of distinguishable orderings is $n!/v!(n - v)!$.

[61] Record at 9–12. Before excuses were granted, each venire contained 60 men, but there was evidence that the proportions of Negroes and whites were not altered by this weeding out. Record at 18. We assume 30-man venires because some of the evidence bore on the venires' composition after weeding and because this is, in this context, a conservative hypothesis: the smaller the panel, the more likely it is to contain a disproportionate percentage of one race.

one of which was 23% Negro. We assume further that Negro males constituted one-quarter of the adult male population.

Applying the binomial distribution previously derived, the probability that not more than five Negroes would appear on any thirty-man venire is:

$$P(v \leq 5) = \sum_{v=0}^{v=5} \binom{30}{v} P^v Q^{30-v}$$

On the assumption that the probability P of selecting a Negro venireman is one-quarter and the probability Q of selecting a white venireman is thus three-quarters the above expression becomes:[62]

$$P(v \leq 5) = \sum_{v=0}^{v=5} \binom{30}{v} (1/4)^v (3/4)^{30-v}$$
$$= 0.20260$$

This result means that about one venire in five should have five or fewer Negroes.

In *Swain*, however, venires with five or fewer Negroes appeared in thirty consecutive cases. The probability of this occurrence, applying the product rule, is $0.2^{30} = 4.63 \times 10^{-21}$. This means that, on the average, only one in more than one hundred million trillion groups each containing thirty venires would consist solely of venires which were not more than 15% Negro. If thirty jury venires were selected at random in Talladega County every day of the year, the daily selection would correspond to the facts in *Swain* only one day, on the average, in thousands of trillions of years.[63]

---

[62] Computed from HARVARD COMPUTATION LABORATORY, TABLES OF THE CUMULATIVE BINOMINAL PROBABILITY DISTRIBUTION 137 (1955). Since the table gives the probability that v is *greater* than or equal to any given value, it is necessary to let r in the table $= v + 1$ and subtract the result from 1.00000 to determine the probability that $v \leq 5$. In the example of the text $v = 5$; $r = v + 1 = 6$; $P = 0.25$. The table shows the value for $P(r \geq 6) = 0.79740$ which when subtracted from 1.00000 yields $P(v \leq 5) = 0.20260$. Statisticians frequently compute the value of $P(v)$ by using approximations such as the Poisson or normal approximations. *See* 1 W. FELLER, AN INTRODUCTION TO PROBABILITY THEORY AND ITS APPLICATIONS 142 (2d ed. 1957).

[63] The assumption that thirty such venires were chosen in a row receives only general support from the record since it is not clear exactly how many were chosen. Record at 10. But our conclusions remain unaffected within a broad variation in numbers. If only twenty venires were chosen instead of thirty, the probability that all would have five or fewer Negroes would be the negligible 1.05×10^{-10}.

The petitioner in *Swain* took the position that the *average* proportion of Negroes on venires was 10 to 15%. Brief for Petitioner at 19. The Court apparently did

If all the *Swain* venires selected since 1953 are included, the total rises to 35, and one venire was 23% Negro. Although the numerical result is not significantly changed by considering this data, we shall do so to illustrate an extension of the method.

We seek the probability of the occurrence of thirty-four venires with five or fewer Negroes and one venire with seven or fewer Negroes. This can be expressed as the sum of the probability of thirty-five venires with five or fewer Negroes, which is $P(v \leq 5)^{35}$, and the probability of thirty-four venires with five or fewer Negroes and one venire with six or seven Negroes. This latter probability, using the binomial distribution, is $35!/34!1!\ P(v \leq 5)^{34} P(6 \leq v \leq 7)$. The sum of these two probabilities is:

$$
\begin{aligned}
P(s) &= P(v \leq 5)^{35} + \frac{35!}{34!\,1!} P(v \leq 5)^{34} P(6 \leq v \leq 7) \\
&= (0.2)^{35} + (35)\,(0.2)^{34}\,(0.3) \\
&= 1.85 \times 10^{-22}
\end{aligned}
$$

This is about twenty-five times smaller than the probability computed without the additional data.

It is instructive to apply these methods to *Cassell v. Texas*[64] (although that case involved jurors and not veniremen) where information concerning the grand juries was more precise than in *Swain.* Of the twenty-one grand juries considered in *Cassell*, a single Negro appeared on seventeen and none on the remaining four. Negroes comprised 6.5% of the eligible population. The Court held that these statistics did not constitute evidence of discrimination since the average number of Negroes on the juries when considered as a group was 6.7%. This did not appear to differ significantly from their proportion in the population of eligibles. Mr. Justice Reed and the three Justices who joined in his opinion did not base their reversal of the conviction on the improbability that a series of juries would be selected at random

---

not accept this interpretation, and the weight of the evidence seems to favor the Court's understanding that 15% was the *maximum* proportion of Negroes. *See* 380 U.S. at 205. Why petitioner's counsel adopted this less favorable view of the evidence is mystifying. Yet even the probability that venires would average 15% Negro is minute. If we again assume the selection of 30 venires with 30 members each out of the 900 veniremen selected, 15% or 135 are taken to have been Negroes. While the exact computation of the probability of this occurrence would be laborious, its smallness is indicated by the fact that the probability of choosing not more than 170 Negroes is less than 0.00001. *See* HARVARD COMPUTATION LABORATORY, *supra* note 62, at 205. The probability of choosing not more than 135 Negroes would be substantially less than this, although larger than that computed on the Court's assumption.

[64] 339 U.S. 282 (1950).

in which the proportion of Negroes never exceeded to any large degree their proportion in the population. On the assumption that $P = 0.065$ and $Q = 0.935$, and using the formulas previously derived, we have $P(v \leq 1) = 0.818$. The probability that not more than one Negro would appear on each of 21 venires (a better distribution than that actually observed since there were four venires without Negroes) is $0.818^{21} = 0.015$. Whether this result[63] supports Justices Frankfurter, Burton, and Minton, who grounded their concurrence on the improbability that no more than one Negro would ever appear on a venire, depends on the statistical test used to evaluate the probabilties.

Our computation of the probabilities in *Swain* was based on the assumption that the selection of veniremen corresponded to a series of Bernoulli trials with a probability $P = \frac{1}{4}$ of selecting a Negro venireman. On this assumption there was apparently only a minute probability that so few Negroes would be selected. Since an event with such an apparently small probability did occur, we are led to reject the basis on which the probabilities were computed, namely, the assumption that venire selection in *Swain* consisted of Bernoulli trials with the probability $P = \frac{1}{4}$ of selecting a Negro venireman. This reasoning is formalized as the following statistical test:

The assumption that the racial result of venire selection corresponds to a series of Bernoulli trials with probability $P = \frac{1}{4}$ of selecting a single Negro venireman is rejected whenever the probability of selecting not more than the observed number of Negroes on venires is less than a certain critical value. For purposes of illustration the critical value $P(v) = 0.05$, the value most commonly used by statisticians, is selected.

Since the probability of selecting not more than the observed number of Negroes on venires in *Swain* was vastly smaller than 0.05, application of the test leads us to reject the hypothesis that the selection of veniremen in *Swain* corresponded to the assumed series of Bernoulli trials. Rejecting the hypothesis that veniremen were selected at random, with the probability of selecting a Negro equal to the proportion of Negroes in the population, does not in itself imply discrimination in the selection process. Technically speaking, all we know is that the assumed Bernoullian model with $P = \frac{1}{4}$ is not consistent with the observed facts. To determine whether there is discrimination we must consider the alternative hypothesis that P does not equal ¼. One fourth of the adult males in Talladega County were Negroes. But if — as the Court was apparently willing to believe — the jury commissioners considered a larger proportion of Negro adult males than white adult males to be unacceptable for jury service, P would be less than ¼. This would only be unconstitutional discrimination if the commissioners applied higher standards for Negroes than for whites or if they excluded from their consideration a substantial number of qualified Negroes.[66] Since the absence of qualified Negroes is the only constitutionally valid reason for accepting a smaller value for P, the qualifications-issue which lurked in the wings is now on stage.

---

[63] The statistical result here resembles that reached by Ulmer, *Supreme Court Behavior in Racial Exclusion Cases: 1935–1960*, 56 Am. Pol. Sci. Rev. 325, 329 (1962).

The statute governing jury selection in Talladega County, which is not untypical, provides that a juror must be a male citizen, generally reputed to be honest and intelligent, esteemed in the community for integrity, good character, and sound judgment, over the age of twenty-one, not a habitual drunkard, able to read English (unless he is a freeholder), and never convicted of an offense involving moral turpitude.[67] The Court pointed out in *Swain* that the method used to find qualified jurors need not be perfect in the sense that every segment of the population receives an equally scrupulous canvass.[68] The nature and range of qualifications required of jurors and the latitude allowed in searching out veniremen combine to make the qualifications issue a substantial problem.

Analysis of this problem requires a reversal of our former procedure. Before, the probability $P(v)$ of observing not more than v Negroes on the venires was calculated from P, the probability of selecting a single Negro venireman. We now seek a value for P which will yield a value for $P(v)$ which is equal to or greater than the critical value 0.05 and would thus by our test result in the acceptance of the Bernoullian hypothesis. In other words, if it is contended that P is really less than ¼, we will inspect other values of P to determine how small it must become before $P(v)$ would satisfy the statistical test. P thus determined is the largest proportion of eligible Negroes to the total eligible population which it is possible to assume and still obtain a probability for the observed results which is not less than the critical value 0.05.

---

[66] Cassell v. Texas, 339 U.S. 282 (1950). *But see* Rabinowitz v. United States, 35 U.S.L.W. 2061 (5th Cir. July 20, 1966) (en banc) (federal jury lists must be "fair cross-section" of community).

[67] ALA. CODE tit. 30, § 21 (1958).

[68] 380 U.S. at 208, 209.

In making these determinations it is mathematically possible to reverse things and to express P as a function of $P(v)$. This, however, involves some technical problems and denies us the use of tables for computing the binomial distribution. It is sufficient for our purposes to compute $P(v)$ for diminishing values of P until the point appears at which $P(v)$ passes the critical value of 0.05. Table I relates differing values of P to $P(v)$ on the assumption that $P(v) = P(v \leqq 5)^{30}$.

TABLE I[69]

Probability of Selecting a Single Negro Venireman P	Probability of Selecting a Single Venire of 30 with 5 or Fewer Negroes $P(v \leqq 5)$	Probability of Selecting 30 Such Venires $P(v \leqq 5)^{30}$
.25	.20260	4.63×10^{-21}
.20	.42751	8.4777×10^{-12}
.15	.71058	3.539×10^{-5}
.12	.85692	0.0097
.11	.89509	0.03598
.10	.92681	0.1002
.05	.99672	0.9053
.02	.99997	0.9911

The table shows that the value $P(v)$ passes from less than 0.05 to more than 0.05 as P passes from 0.11 to 0.10. Consequently, in order to obtain a value of $P(v)$ which meets the test of the Bernoullian hypothesis P would have to have a value less than 0.11. For simplicity it is assumed that the critical value is $P = 0.10$; this is accurate enough for present purposes. In other words, we cannot account for the observed number of Negroes on venires unless we assume that eligible Negroes constitute 1/10 or less of the total eligible population.

Since Negroes comprise approximately 25% of the adult male population, there are three times as many adult male whites as adult male Negroes. On our assumption that no more than 10% of the eligibles are Negroes, there must be at least nine times as many eligible whites as eligible Negroes. This means that, to satisfy the statistical test established for the Bernoullian hypothesis, it must be assumed that the proportion of adult male whites qualifying is at least three times as large as the proportion of adult male Negroes qualifying. For example, if all the whites qualify, then not more than one-third of the Negroes may qualify; if 75% of the whites qualify, then not more than 25% of the Negroes may qualify.

---

[69] This table was computed as described in note 62 *supra*.

Qualifications for jury service may be divided roughly into objective and subjective factors. The most important objective factor is literacy; among the subjective factors are a reputation for honesty and intelligence, community esteem, and so forth. The Court should look first at census and other data to determine the impact of the objective factors. If these do not account for the three-to-one qualification rate in favor of whites, then the effect of the subjective factors must be considered.

The Census Bureau has not reported how many Negroes in Talladega County have completed six years of school (and are thus presumably literate[70]), but the statewide average census figures show that approximately 48% of adult male Negroes and 83% of adult male whites have had such schooling.[71] If this average applies to Talladega County, the greater rate of illiteracy among Negroes cannot by itself explain their absence from the venires, since the comparative literacy ratio is not three-to-one but less than two-to-one. Assuming that literacy is the only significant objective factor,[72] the three-to-one qualification rate must be owing in part to the application of the subjective criteria.

An eligible venireman must be literate and must also satisfy the subjective criteria. We must now determine the proportion of the adult male literates of each race qualifying on subjective grounds. Since there are three times as many whites as adult Negroes, and since 83% of the whites and 48% of the Negroes are literate (using the sixth grade schooling test), the ratio of adult literate whites to adult literate Negroes is:

$$\frac{3 \times 83\%}{1 \times 48\%} = 5.2$$

---

[70] This is the standard for presumptive literacy used in the Voting Rights Act of 1965. 42 U.S.C. §§ 1971(c), 1973b(e) (Supp. I, 1965).

[71] DEPARTMENT OF COMMERCE, BUREAU OF CENSUS, U.S. CENSUS OF POPULATION: 1960, vol. 1, pt. 2 (Alabama). Talladega County Negroes over the age of twenty-five are more literate than adult Negroes in the state as a whole, since their median schooling is 6.5 years. *Id.* at 2–201, 2–224.

[72] This implies that conviction of an offense involving moral turpitude and habitual drunkenness, the two other objective factors mentioned by the Alabama statute, do not markedly affect the relative proportions of Negroes and whites qualifying.

Thus, when adult males emerge from the literacy test there are about five times as many whites as Negroes. But we know that, after the literacy and subjective tests, there are at least nine times as many whites as Negroes. We know, therefore, that whites are almost twice ($9/5$) as successful as Negroes in getting through the subjective tests. In other words, the proportion of literate Negroes qualifying on subjective grounds is $5/9$ of the corresponding proportion of white literates.[73]

The qualifications issue has now been refined to the point where the problems of legal policy are exposed for the Court to deal with unencumbered by subsidiary statistical questions. The principal legal question is whether a state, by imposing subjective criteria for jury service or by its method of administering the system, may act in such a way that the rate of qualification among literate Negroes is substantially smaller (in our case by almost half) than the rate among literate whites. Several resolutions are possible. The Court might hold such a result unconstitutional per se. It might view the result as prima facie evidence of discrimination which the defendant would be at liberty to rebut. Or it might require the petitioner to present evidence of specific malfeasance. Whatever the resolution, these alternatives present problems of basic importance to this branch of the law. The Court did not reach those problems in *Swain* because of its inability to assess the significance of statistical data without mathematical tools.

---

[73] This line of reasoning may be expressed algebraically as follows: If EN/LN equals the proportion of eligible Negroes in the literate Negro population and EW/LW equals the proportion of eligible whites in the literate white population, we are seeking to determine the number a such that $EN/LN = (a)(EW/LW)$. Transposing, $a = (EN/EW)(LW/LN)$. We have determined that the proportion of eligible Negroes to eligible whites (EN/EW) must be no greater than $1/9$. Thus, at most, $EN/EW = 1/9$. We have also determined that the proportion of literate whites to literate Negroes (LW/LN) is approximately 5. Combining the two: $a = (EN/EW)(LW/LN) = (1/9)(5) = 5/9$, the result given in the text.

When we examined the representation of Negroes on venires without considering the qualifications issue, the statistics were so extreme that it made little difference which assumptions or methods were used. This is no longer true when dealing with the qualifications issue. If twenty rather than thirty venires were assumed chosen consecutively, P would be closer to 0.12. The simultaneous use of other assumptions favorable to the state could raise this figure even higher. On the other hand, we have omitted information which cuts the other way, *e.g.*, the testimony that during this period there were three or four grand juries without *any* Negroes. Record on Appeal at 10. It is not necessary for our purposes to work out the numerical consequences of these possibilities. We do note that our results are sensitive to our assumptions, and that if the Court had reached the qualifications issue, it might reasonably have decided that more accurate information was essential for decision.

It is appropriate to consider the significance of using the critical value 0.05 in the foregoing analysis. The size of the critical value affects the probability of error. When will a test with a critical factor of 0.05 lead to wrong results? In considering this question, statisticians usually divide error into two types. Type I error is the error of rejecting the hypothesis of random selection when it is true. When the critical value 0.05 is used, this type of error will occur, on the average, not more than one time in twenty — on the one occasion in twenty when a series of Bernoulli trials yields the observed distribution. Only the state has standing to complain of type I error since it involves a finding of discrimination when there is none.

Type II error is the error of accepting the random selection hypothesis when it is false. The extent of such error is more difficult to determine. However, only the defendant has standing to complain of type II error since it involves a finding of non-discrimination when in fact there is discrimination. When a court determines with the selected test that discrimination is present, the only error it need consider is type I error. It is only when no discrimination is found that type II error need be considered.

It should be noted that on the facts in *Swain* a substantially smaller critical value (and thus substantially greater accuracy) does not change the results very much. Thus if the critical value were 0.01 then P would equal about 0.12[74] and the rate of disqualification for the subjective factors would be recomputed as 6.1%. This is not materially different insofar as the legal issues are concerned from the 5% rate which was computed using the critical value 0.05.

Are a critical value of 0.05 and a type I error of one-in-twenty too large? This is a legal issue for which there can be no firm answer. The Court must weigh the risk of intruding on a state's selection process without cause against the risk of upholding venires which were in fact chosen discriminatorily. An 0.05 test, however, seems consistent with another case decided by the Court. In *Avery v. Georgia*[75] defendant was convicted by a jury selected from a panel of sixty veniremen. The veniremen were chosen from a box containing tickets with names of the persons on the jury roll — yellow tickets for Negroes and white tickets for whites. Five percent of the tickets in the box were yellow, but not a single yellow ticket was selected for the venire of sixty.

---

[74] *See* Table I, p. 361 *supra*.

[75] 345 U.S. 559 (1953).

The Court held that the all-white venire and the opportunity to discriminate presented by the yellow tickets constituted a prima facie case of discrimination. Mr. Justice Frankfurter, concurring, concentrated on the absence of Negroes from the venire: "The mind of justice, not merely its eyes, would have to be blind to attribute such an occurrence to mere fortuity."[76]

The statistical case is hardly that conclusive. The probability of selecting a white is $Q = 0.95$. Applying the multiplication rule, the probability that sixty whites would be selected is $Q(w = 60) = (0.95)^{60} = 0.046$. This probability is enormously greater than that computed in *Swain*. The Court's conclusion in *Avery* is acceptable under our test only if a critical value of the order of 0.05 is viewed as sufficiently small. It may have been, however, that the Court's intuitive evaluation of the probabilities was influenced by its knowledge that the colored ticket system furnished a way to discriminate and suggested an intent to do so.

B. *An Analysis of the Selection of Grand Jurors*

We now consider the issue which the Court would have faced in *Barksdale v. Louisiana*[77] had it granted certiorari, namely, whether a given distribution of Negroes on grand juries is consistent with random selection.

For problems of this type statisticians frequently use a goodness of fit test known as the chi-square (χ^2) test.[78] In essence the chi-square test as applied to this problem is a measure of the probability that the observed distribution of Negroes on grand juries would differ to the degree it does from the distribution which would be most probable if the selection of jurors were truly random. By distribution, I mean the proportions of grand juries with exactly one Negro, exactly two Negroes, exactly three Negroes, and so forth. To construct the test one must assign a measure to the difference between the observed and expected distributions, and then determine the probability that a distribution with a deviation at least as large as the observed deviation will occur. The less probable such a deviation, the more we suspect our hypothesis. The theory of the chi-square test is simple, although actual computation of the probabilities leads to some technical problems.

---

[76] *Id.* at 564.

[77] 247 La. 198, 170 So. 2d 374 (1964), *cert. denied*, 382 U.S. 921 (1965).

[78] *See* Robinson, *Bias, Probability, and Trial by Jury*, 15 AM. SOCIOLOGICAL REV. 73 (1950). For a nontechnical explanation of χ^2 see J. HODGES & E. LEHMANN, BASIC CONCEPTS OF PROBABILITY AND STATISTICS 285–88 (1964). More extensive discussions may be found in R. FISHER, STATISTICAL METHODS FOR RESEARCH WORKERS 78–113 (13th ed. rev. 1958); Fisher, *The Conditions Under Which* χ^2 *Measures the Discrepancy Between Observation and Hypothesis*, 87 J. ROYAL STATISTICAL SOC'Y 442 (1924). Tests based on other factors such as the occurrence of "runs" are also possible. *See* David, *A* χ^2 *'Smooth' Test for Goodness of Fit*, 34 BIOMETRIKA 299 (1947); Swed and Eisenhart, *Tables for Testing Randomness of Grouping in a Sequence of Alternatives*, 14 THE ANNALS OF MATHEMATICAL STATISTICS 66 (1943).

If $P(v)$ is the probability that a jury with exactly v Negroes would be chosen at random, and if N juries are chosen, the number of juries with v Negroes most likely to be selected is $NP(v)$. This is called the expected number of juries with v Negroes, and is abbreviated as $E(v)$.

We are interested in the difference between the expected and observed number of juries for every proportion of Negroes from 0% to 100%. This difference is $O(v) - E(v)$, where the observed number of juries with v Negroes is $O(v)$ and the expected number of such juries is $E(v)$. If, for example, the probability of selecting a jury with two Negroes is $P\ (v = 2) = \frac{1}{4}$ and four juries were chosen we would expect that one of the juries would have two Negroes since $E(v = 2) = NP(v = 2) = 4 \times \frac{1}{4} = 1$. If in an actual drawing of four juries none had two Negroes, then the difference between the number observed and the number expected would be $(0 - 1) = -1$.

In order to obtain a total measure for the deviation from the expected number of Negroes on juries the difference between the observed and expected values is summed up for each v from $v = 0$ to $v = 12$ (the size of each jury). In other words we compare the difference between the observed and expected number of juries with 0, 1, 2 . . . 12 Negroes, add these differences, and obtain a single number which measures the totality of the differences. Since the observed values may be greater or less than the expected values, these differences would in some cases be negative (as in the example) and thus reduce the sum. This would be misleading because in this context it is irrelevant which is greater, the expected or observed frequencies; the only significant fact is that they are different. For this reason, mathematicians usually consider the square of each difference, which is always positive. Finally, in order to make the value of the differences proportional to their expected frequencies each squared difference is divided by its associated expected frequency.[79] We thus obtain the following expression for χ^2 (chi-square) as the sum of the squares of the differences between the observed and expected frequencies divided by the expected frequency:

---

[79] If, for a particular v, $O(v) = 231$ and $E(v) = 230$, the difference between them would be 1. But this is a much better agreement between observation and

$$\chi^2 = \sum_{v=0}^{v=12} \frac{(O(v) - E(v))^2}{E(v)}$$

Note that χ^2 increases as the difference between the observed and expected frequencies increases and decreases to zero as the sum of the differences between the observed and expected frequencies diminishes.

Given a particular distribution of Negroes on juries, the issue presented is: what is the probability of obtaining a distribution which deviates to this or a greater extent from the expected distribution? In terms of the χ^2 terminology this is equivalent to asking the probability of a distribution with an associated χ^2 value which equals or exceeds the χ^2 of the observed distribution. The answer is obtained by computing the value of χ^2 for every possible distribution, determining the distributions with values of χ^2 which are equal to or greater than the given χ^2, computing the probability of the occurrence of each such distribution (using the probability formulas previously derived), and then summing up these probabilities. The resulting sum is the probability of the occurrence of a distribution with a χ^2 which equals or exceeds the χ^2 of the observed distribution. This probability is called the χ^2 probability.

The formula for χ^2 rests on the assumption of a series of Bernoulli trials with constant probability P of selecting a Negro. If there is an extremely small probability of a distribution with a χ^2 value equal to or greater than the observed χ^2 value, doubt is naturally cast upon the assumption. The χ^2 probability should then be computed using other values of P. If this leads to a sufficiently large χ^2 probability, the value of P used to compute χ^2 should then be compared in the same way as in the test previously described with the evidence of qualifications of the Negro community.

The χ^2 test differs in an important respect from the previous test. Since in the previous test we sought to determine the probability that so few Negroes would be chosen, it was always possible

---

expectation than if $O(v) = 2$ and $E(v) = 1$. Dividing by $E(v)$ brings out this proportionality factor. Thus in the first example $\chi^2 = \frac{1}{230}$ but in the second $\chi^2 = 1$.

to find a value for P small enough to account for the observed racial population of the venires. If very few Negroes appeared on venires, P would have to be assumed to be very small, and thus inconsistent with the actual qualifications of the Negro community. Still, in every case some value of P could be found which made the observations consistent with the hypothesis of randomness. But since, in dealing with the χ^2 test, we are determining the probability of the appearance of a particular *distribution* and not just a maximum number of Negroes, we cannot in all cases account for the observations merely by reducing the value of P. The value of P which maximizes the χ^2 probability may still lead to a value too small to be consistent with the Bernoullian hypothesis.

This property becomes important since it enables us to apply the χ^2 test to distributions when P is unknown. In such cases we assume that value for P which will lead to the maximum χ^2 probability. If the χ^2 probability so computed is still too small to be consistent with the Bernoullian hypothesis, then that hypothesis must be rejected since no value of P would lead to its acceptance. In other words, no matter what we are willing to assume about Negro qualifications, if we find, say, that every grand jury contains one Negro, neither more nor less, we may be able to conclude that selection has not been random.

These observations may be formalized as the following test: *The assumption that the racial result of grand jury selection corresponds to a series of Bernoulli trials is rejected whenever a distribution appears for which the maximum χ^2 probability is less than a selected critical value.* For purposes of illustration the critical value $P(v) = 0.05$ is selected. The same analysis of type I and II error may be applied to the χ^2 test as to the previous test.

We now apply the χ^2 test to the statistics presented to the Court in the petition for certiorari in *Barksdale.* In Orleans Parish there were nine grand juries selected between September 1958 and September 1962, when the grand jury which indicted Barksdale was impaneled. There were two Negroes on eight of these juries and one on the ninth. Each grand jury had twelve members. Negroes comprised one-third of the adult male population of the parish, but the Supreme Court of Louisiana held that the smaller number of Negroes on grand juries was explained by their lower level of literacy and by the fact that more Negroes than whites requested excuses because of economic hardship. There was no evidence as to the number of Negroes and whites requesting excuses. For this reason it is assumed that P is some unknown value less than $\frac{1}{3}$.

It will be recalled that χ^2 is computed from the differences between the observed and expected numbers of Negroes. Determining the observed numbers of Negroes $O(v)$ is easy. Since there was one jury with one Negro and eight juries with two Negroes, $O(1) = 1$; $O(2) = 8$; $O(v) = 0$ (for all other v). Determining the expected numbers of Negroes $E(v)$ is more difficult. By the formulas previously derived, for each v:

$$E(v) = NP(v) = \frac{9(12!)}{(v)!(12-v)!} P^v Q^{12-v}$$

In this equation, 9 is the value of N, the total number of grand juries impaneled. In order to determine $E(v)$ it is necessary to determine P. Since P is treated as unknown, we select a value for P which will maximize the χ^2 probability. This occurs when the value of χ^2 is at a minimum, since at that point there are the greatest number of distributions with χ^2 values which equal or exceed the χ^2 of the observed distribution. In other words, we are looking for the value of P which will yield the expected distribution most in accord with the observed distribution. The value of P which will lead to minimum χ^2 can be determined in most cases by using differential calculus.[80]

Computing the χ^2 probability on the facts here is a forbidding task. The χ^2 value for each possible distribution of Negroes and whites on juries must be computed and the probability determined for each distribution with a χ^2 value which is not less than the χ^2 of the observed distribution. If both the number of jurors and the number of juries is large the number of possible distributions of Negroes becomes extremely large, and the calculation of χ^2 and the associated probabilities becomes too long even for modern computers operating within a reasonable span of time. Fortunately, as these numbers become large it is possible to approximate the χ^2 probability by means of an approximation function known as the χ^2 distribution, whose values are widely available in tables.[81]

The approximation function cannot be used with any confidence, however, to compute the χ^2 probability in *Barksdale*. The sample is too small for this purpose. The facts in *Barksdale*, however, permit a simplification which avoids the need for a computer. Since the number of Negro jurors remained almost constant over the nine juries, a simple test can be constructed based on a determination of the probability that the numbers of Negroes appearing on juries would have been thus constant if the selection of jurors was random. To test this issue using a version of the χ^2 approach, we begin by asking how many possible distributions of seventeen Negroes on the twelve juries have a variation not greater than the variation of the observed distribution. The answer is clear: the observed distribution (two Negroes on each jury save one which had one) is the most constant possible; any other distribution of seventeen would show a greater variation. Thus, out of all possible distributions of seventeen Negroes, the probability of observing a distribution which varied as little as that observed is simply the proportion which the probability of the observed distribution bears to the total probability that seventeen Negroes would have been selected regardless of distribution.[82]

---

[80] In the instant case χ^2 would be a minimum when $P = 0.165$, which is close to $\frac{1}{6}$ or the average proportion of Negroes on the nine juries.

[81] *E.g.*, P. Hoel, *supra* note 56, app. 2.

The computation of probabilities of this sort has already been discussed and no new principles are involved. There were 17 Negroes chosen out of 108 jurors. The assumption most favorable to the state (the one that leads to the largest probability) is that $P = 17/108$. Using this assumption and the binomial distribution previously derived, the probability of selecting a single jury with two Negroes is 0.295. Upon the same assumption, the probability of selecting a jury with only one Negro is 0.287. The probability that eight juries would have two Negroes each and the ninth would have one Negro is, by the binomial distribution:

$$\binom{9}{1} (0.295)^8 (0.287) = 0.000148$$

---

[82] The method used here is a special example of a test of the possibility, not present in most applications of statistical methods, that the selecting agents might manipulate selection to produce a result which closely approximated or equaled the expected distribution. Testing for such practices involves a computation of the probability that the observed distribution would differ to such a small extent from the expected distribution. If the χ^2 test were used, the probability of each distribution with a χ^2 equal to or less than the χ^2 of the observed distribution would have to be summed, and if this sum were too small the hypothesis of random selection would be rejected. For example, the probability of precisely the expected distribution occurring is usually so small that, if it were observed, statisticians would reject the random selection hypothesis.

A really sophisticated group of jury commissioners bent on discrimination would have to select groups of juries which were neither too close nor too far from the expected. Against such educated malevolence the chi-square test offers no defense.

This is the probability of the occurrence of the observed distribution out of all possible distributions of Negroes on juries. The total probability of selecting seventeen Negroes regardless of distribution is:

$$P(v = 17) = \binom{108}{17}\left(\frac{17}{108}\right)^{17}\left(\frac{91}{108}\right)^{91} = 0.104$$

There is thus approximately one chance in ten that seventeen Negroes would be selected. Dividing the two probabilities yields:

$$\frac{0.000148}{0.104} = 0.0014$$

The conclusion is that, on the assumption most favorable to the state, there is approximately one chance out of a thousand that a distribution of seventeen Negroes on twelve grand juries would show a variation as small as that observed. Since this is much smaller than the critical value 0.05 we are led to reject the Bernoullian hypothesis.[83]

The significance of this rejection must now be considered. Rejection of the Bernoullian model of the selection process can be treated as implying the existence of discrimination only if this model would accurately reflect the racial results of jury selection in a system free from discrimination. It is important to observe that whether this is so is not a logical but an empirical question. It can be resolved only by charting the results of a system which is free from discrimination but which otherwise operates in precisely the same way as the system under experiment. But although scientific verification is lacking, the assumption we have used seems plausible.

---

[83] In applying a version of the χ^2 method to the facts in *Barksdale* we sought to measure the probability that the number of Negroes on grand juries would reflect almost undeviatingly their proportion in the population of eligibles. An alternative test of the same issue could be constructed by determining the probability under a random selection system that no juries would have had fewer Negroes and no juries would have had a greater number of Negroes. The probability that no jury would have more than two Negroes and at least one would have fewer than two is $P(v \leq 2)^9$, that is the probability that all nine juries would have had two or fewer Negroes, less $P(v = 2)^9$, the probability that all nine juries would have two. Upon the assumption that $P = 17/108$ this probability is approximately 0.046. There are thus slightly less than five chances in one hundred that a greater number of Negroes would not have appeared on some juries if the selection were random.

The other half of this proposition is that there should have been some juries with fewer Negroes than were observed. The probability that every jury would have at least one Negro and that at least eight would have two or more is $9P(v = 1)(1 - P(v \leq 1))^8$, or the probability of one jury having a single Negro and the remainder having two or more, plus $(1 - P(v \leq 1))^9$, or the probability of all juries having two or more Negroes. Upon the same assumption concerning P this probability is approximately 0.043. There are thus slightly less than five chances in one hundred that a smaller number of Negroes would not have appeared on some juries if the selection were random.

The probability of the joint occurrence of both events is the product of the probabilities of each. This is 0.046 × 0.043 = 0.0019. Thus in only two chances out of a thousand would the number of Negro jurors adhere so closely to their proportion in the eligible population. This is similar to the result obtained by the method used in the text.

If the observations are of venires, then to reject the Bernoullian hypothesis implies that P must vary from venire to venire, or from the selection of one venireman to the selection of the next. Since, within each system now in use, the method of selection is supposed to remain the same for each venire and venireman, the only reason for variation would seem to be intentional intervention: when the officials charged with selecting veniremen choose to exclude Negroes, $P = 0$ and $Q = 1$. It is this reasoning that makes plausible the assumption that an honest selection process consists of a series of Bernoulli trials and that a non-Bernoullian process is discriminatory.

Does this also apply to the grand jury? In *Barksdale* it was claimed by the state that the relative absence of Negroes from grand juries could be explained, *inter alia*, by their more frequent requests to be excused because of economic hardship. Assuming such requests occurred, can it be said that selecting grand jurors is, in the absence of discrimination, a series of Bernoulli trials? There appears no reason to treat the selection of grand jurors differently from the selection of veniremen. The existence of an excuse based on economic grounds might affect the number of available Negroes, but within the pool of eligibles the selection of grand jurors must still be made without consideration of race. Poverty can explain why only one-sixth as many Negroes as whites are left in the pool, but not why precisely one-sixth of almost every grand jury selected is Negro. Since economic hardship and illiteracy operate in similar ways as barriers to jury service it appears reasonable to extend the assumption that the selection of venires is Bernoullian to the grand jury.

Whether this assumption may be extended to the petit jury is much more doubtful. If it is constitutionally permissible for the prosecutor to strike any Negro from the jury in a capital case where the defendant is a Negro, or to follow similar racial rules, the selection process for the petit jury is not Bernoullian because on the occasions when Negroes are excluded $P = 0$. On the other hand, if exclusionary rules are not constitutionally permissible and each juror must be considered on his merits, the case for the Bernoullian hypothesis is stronger since the determination to strike a juror must be more particularized and less dependent on race. But even if this is the law, it is still possible that *P* will be smaller in any case where the prosecutor might reasonably expect a greater possibility of bias on the part of a Negro juror. If that is so, the selection of petit jurors for a series of juries would not consist of Bernoulli trials. In short, it seems fair to extend the logic of the Court's decisions in the venire cases to grand juries, but it is doubtful that the same methods can be applied to petit juries.

Barksdale involved grand juries, so the Bernoullian hypothesis should apply. Yet we have seen that, even assuming that the proportion of Negroes on grand juries in Orleans Parish was the same as their proportion in the population of eligibles, the fact that it was almost unvaryingly so makes the distribution inconsistent with the hypothesis of random selection. Since on the assumption most favorable to the state the distribution test leads to a rejection of the hypothesis of random selection, we must conclude that no argument based on the presence or absence of qualifications in the Negro community will explain the observed distribution. It is overwhelmingly unlikely that the *Barksdale* facts could have resulted from the evenhanded use of selection criteria unfavorable to Negroes or from sloppy, haphazard administration. Within the confidence limits of the test, the conclusion seems inescapable that race was considered in selecting grand jurors.

IV

"The Civil War Amendments," Mr. Justice Frankfurter wrote, concurring in *Cassell v. Texas*, "did not turn matters that are inherently incommensurable into mere matters of arithmetic."[84] To the reader who has faithfully threaded his way through this article it should be apparent that the mathematical methods developed here do not supplant the necessity for legal judgments. The purpose and effect of the mathematical methods is to draw implications from the data on which the legal judgments are founded. If the consequence of such analysis is a legal decision that discrimination exists where uninformed intuition would have reached a different conclusion, the decision was compelled, not by mathematics, but by the consistent application of legal principles to new information.

---

[84] 339 U.S. 282, 291 (1950).

A basic legal principle in the jury discrimination cases is that the selection of an improbably small number of Negroes is evidence of discrimination. This principle, which links a finding of discriminaton to a determination of probabilities, opens the door to the use of statistical analysis in these cases. The mathematical methods described here have been used to calculate the probabilities which the law has established as relevant for determining the existence of discrimination.

The second legal principle controlling these cases is that a disparity between the proportion of Negroes on venires and in the population generally is evidence of the improbability of random selection. Using this assumption we have constructed a statistical test for randomness and applied it to the number of Negroes appearing on the jury venires described by the record in *Swain v. Alabama.* The application of the test led to the conclusion that the selection process was not random with respect to race if the entire adult male population of both races was considered eligible. The results could not be justified by any lack of qualifications for jury service among Negroes unless it was assumed that the rate of Negro qualification with respect to such subjective factors as lack of integrity was little more than half the rate of white qualification.

The failure to select a greater number of Negroes for venires is not the only fact relevant to a claim of discrimination. A skewed distribution may also constitute such evidence. In order to test for skewness we used a goodness-of-fit test known as the chi-square test to analyse the distribution of Negroes on grand juries in *Barksdale v. Louisiana* and concluded that race was considered in selecting grand jurors in Orleans Parish.

The particular statistical tests which produced these results are not by any means the only possible tests of these matters. Other tests with different characteristics are possible and have been used in treating cognate problems. Unfortunately, an adequate discussion of the criteria for choosing tests would lead us into the vasty deep, and the reader may feel that the waters here have been deep enough already. It must suffice to say that the choice of tests depends on the nature of the suspected discrimination and consequently presents a delicately mixed problem of mathematics, fact, and law.

It may be felt, however, that all mathematical tests are inappropriate for jury discrimination cases because such methods are too technical, elaborate, and fine-grained. Perhaps judicial intuition is sufficient to decide what the Constitution requires and only the intuitively obvious cases should be deemed appropriate for judicial correction. This feeling may be supported by the argument that the Court has never required jury selection to be mathematically perfect in the sense that every potential venireman be canvassed for service; consequently, it may be said that the proposed mathematical technique is misleading because it applies exact methods to an inexact system. This charge is worth considering because it can be leveled at virtually every attempt to apply methods from the exact sciences to legal problems: in each case the legal rules will invariably reflect the human imperfections of our institutions, but the methods used to analyze them will not, seemingly, allow for such imperfections.

In the present context this does not appear to be a significant problem. Any statistical method of analyzing data tells us the attributes of the system under examination but does not tell us the causes of those attributes, though it may be able to rule out some conceivable causes. Thus in *Swain* we found that in order to sustain the observations it must be assumed that the rate of qualification of Negroes for jury service was substantially smaller than the qualification rate for whites. We assumed that this occurred because the commissioners disqualified a proportionately greater number of Negroes, but it may be that the commissioners only innocently canvassed proportionately fewer literate Negroes than literate whites.

In this and other cases, however, the cause of the result may be legally irrelevant. The Supreme Court has held that an innocent failure to canvass any Negroes at all was not an acceptable excuse for a result which was, in effect, the equivalent of discriminatory selection.[85] In at least some cases, then, a system whose results appear to be discriminatory may be treated as such even though the commissioners convincingly profess good faith. Statistical techniques can be used to identify such cases. Even when a case is not extreme, so that the courts are willing to hear a defense of good faith, mathematical methods may be able, as in the *Barksdale* case, to provide a test of that defense.

The acceptance of mathematical methods undoubtedly implies a certain yielding by judges of their freedom of decision. In *Swain* the mathematical case appears overwhelming, and it leads to a conclusion contrary to that reached by the Court. This consequence of mathematical methods may be a source of judicial reluctance to entrust decisions to them and may tempt the Court to continue the familiar practice of casting legal rules in broad discretionary terms to ensure that legal principles will not embarrass with results that run against the grain of judicial intuition.

---

[85] Cassell v Texas, 339 U.S. 282, 287–90 (1950).

In the developing stages of a legal rule, a reservation of discretion may be the better part of judicial valor, but where principles have become established over a long period the desire to reserve an unmeasured power over their consequences is not an appealing motive for rejecting objective methods. In the sensitive area of the law we have been considering, where cases carry the freight of broad social conflicts, the authority of decisions will not be enhanced by methods which depend unnecessarily on personal and intuitive judgments. As Learned Hand wrote, the authority and immunity of a judge rest "upon the assumption that he speaks with the mouth of others."[86] The novelty of the methods proposed in this article may make them seem strange to the lawyer or the judge, but the analysis provided by statistical theory should, in the long run, make the decisions in the jury discrimination cases a more natural and inevitable consequence of the data which determine them.

Every successful use of technical information by the law has had to travel the path from strangeness to indispensability. The importance of objectivity and consistency should persuade us to follow legal principles in these cases even when their logic leads to ground made less certain by the absence of the assurance — and it is sometimes a false assurance — which intuition brings to judgment.

---

[86] Hand, *Mr. Justice Cardozo*, 52 Harv. L. Rev. 361 (1939).

CHAPTER FIVE

CLASS NOTES

An exercise for applying the ideas about statistical inference contained in the readings is fairly clearcut. We can use the data generated by the previous exercise concerned with sampling, based on our hypthetical town.

We will assume that we have drawn a random sample of 40 households and that we have calculated the percentage of households whose head is Negro.

What estimate can we make of the percentage of Negro households in the population from which the sample is drawn?

What estimate can we make of the average annual income in the population, both white and Negro, based on the information contained in our sample?

As the readings indicate, statistical inference involves not only estimates of population means and percentages; it also involves the testing of null hypotheses about the relationship between characteristics. For example:

We can select a random sample of 100 cases. We can then divide the sample approximately in half, with regard to education, dichotomizing our households into "high" and "low." These two groups can each be considered a random sample of "high" and "low" educational groups in the population.

We can then calculate the average annual income of each of the two groups. The problem, then, is this: We can observe a difference in our samples in average income between those categorized as "high" and "low" with respect to education. Does this difference between the two sample groups represent a difference that would be found in the population from which they are drawn or is it a result of sampling vagaries? The solution to the problem can be found in the selection from Mueller and Schuessler.

CHAPTER SIX

MEASURES OF ASSOCIATION

I

The idea of correlation or association lies at the heart of all scientific research. Without a thorough understanding of that idea, it is impossible to assess the utility or the validity of experimental findings, survey design, and so on.

As John Mueller and Karl Schuessler point out in the first selection of this chapter, the attempt to understand the idea of association involves three objectives. First, we wish to identify the variables which are linked to the phenomenon that interests us. Second, we wish to uncover the pattern of that linkage. And third, we wish to measure the strength of that linkage. In a sense, of course, we are all informal statisticians, as these authors indicate, and we are always using the idea of association or correlation as we try to make sense out of our world. But if we are to be at all scientific or rigorous, we must go beyond this informal sort of thinking and tie down our concepts more precisely.

The first type of association discussed is that existing between two attributes. In an earlier chapter we pointed out that an attribute is a characteristic which is either present or absent--like a light bulb, it is either on or off and its presence is not a matter of degree; and we can use the symbols discussed by Morris Cohen and Ernest Nagel, in a later selection in this chapter, and speak of a characteristic which is either present (X) or absent ($\overline{X}$).

When we ask, then, if X is related to Y, we can set forth our finding in what is called a 2 x 2 table: one row indicates when Y is present, another row indicates when Y is absent; one column indicates when X is present, another column indicates when X is absent. (See Table 1.)

TABLE I

	X (X Present)	$\overline{X}$ (X Absent)
Y (Y Present)		
$\overline{Y}$ (Y Absent		

Each cell in the table gives the frequency of the joint occurrence of the four possibilities: XY, $\overline{X}Y$, $X\overline{Y}$, and $\overline{X}\overline{Y}$. (In the readings we have included a discussion of only one measure of association, namely Yule's coefficient of association or Q, for we think it provides most of the essential logical steps when we decide whether two characteristics are related to one another.)

In the next part of the selection by Mueller and Schuessler, the authors turn to the concept of correlation between variables. It is important for the student to realize that the underlying logic is very much the same as that used in analysing the association between characteristics, although there are the added complexities of dealing with variables rather than attributes. This can be seen most easily by examining Figure 10.2.3 in the selected readings. If, in this illustration, we consider -50 to 0 on the X axis as a situation where X is absent, and 0 to 50 as a situation where X is present; and if we consider 0 to 50 on the Y axis as a situation where Y is absent, and 50 to 100 a situation where Y is present; it is then possible to convert the scatter diagram of Figure 10.2.3 into a 2 x 2 table. We have, in effect, treated variables as if they were attributes by dichotomizing them--a very common device discussed by Mueller and Schuessler.

But all of this should be clear to the student in the following selection which is exceptionally lucid. The reading requires very little in the way of mathematical background and yet if the student grasps the essential points he will have grasped one of the most fundamental concepts in scientific inquiry.

I I

Now it is obvious that if one factor is found to be correlated or associated with another it is no proof that one causes another. Yet there is an important logical connection between correlation and causation which must be examined. A search for causes is central to much of the research in the social sciences and we cannot avoid the difficulties that are involved.

The law, of course, has also long had a concern with causation, perhaps most notably in the field of torts. The view of the law, however, is rather different from that of science. In fact, according to Harper and James in the first selection, when the law turns its attention to the problem of causation there is "a widely recognized confusion." In their discussion the authors claim that their presentation is not a metaphysical one but "an ordinary, matter-of-fact inquiry into the existence or non-existence of a causal relation as laymen would view it, and apparently they hope to avoid the ambiguities that plague the subject. But they too, we think, are inevitably caught up in the horrendous difficulties of trying to equate human responsibility and causation in the complex, multi-faceted flow of events.

Professors Hart and Honore, in the selection of their book, Causation in the Law, quite wisely point out that the law's concept of causality is really a metaphor which may obscure empirical reality, as well as reveal it. And as long as the law continues to pursue its chosen task of singling out the voluntary acts of individuals as the cause of particular events, it seems very likely that the law will continue to be at odds with the scientific viewpoint.

The reading from the book by Professors Cohen and Nagel, Logic and the Scientific Method, summarizes much of the current thinking about causation in a variety of scientific fields. They point out that "common sense" notions of cause and effect are heavily dependent on a model of human behavior in which individuals deliberately intervene in the flow of events to effect a certain purpose--and the model is not very good when applied to the non-human world and is of doubtful worth in the world of human behavior as well. In fact, they argue that "In the light of the great variety in the kinds of specific orders and factors which may be the objects of an inquiry, it may seem preposterous to believe that any general rules can be stated which will enable us to find satisfactory answers to all possible problems." The search for causes thus seems to have an indeterminate and uncertain quality--which may be disturbing for the law in its desire for certainty and personal accountability but which is familiar enough in science.

The fact that it is so difficult to establish cause and effect relationships and that the concept itself is variable in its meaning has led some philosophers of science to urge that we abandon the idea of causation altogether. We would simply deal with correlations, recognizing that correlation and causation are not the same thing. It is doubtful if many scientists would be willing to go so far, but the problems in dealing with the idea of causation do impose a fair amount of caution in scientific inquiry.

First, scientists accept the fact that there are generally many factors which contribute to a particular effect and no one cause can be singled out as the cause.

Second, cause and effect represent a statistical probability of the joint occurrence of two or more events rather than an invariant relationship.

Third, there are many different levels of discourse and many different ways of organizing data. What is examined as a cause in one instance of scientific inquiry may be largely ignored in another.

Fourth, in the examination of social behavior, experimentation is difficult or simply ruled out on ethical grounds. As a consequence, we are seldom able to isolate the influence of one variable and our assertions about causation must be patched together with many untested assumptions.

In short, the evidence that we can accumulate about cause and effect in social behavior is weak and uncertain. But it is far better than no evidence at all.

EXPLAINED VARIATION

In the following selection by Mueller and Schuessler, the authors discuss the measurement of relationships both between attributes and between variables. (They speak of the principle of joint occurrence and the principle of covariation.)

It is the latter--the relationship between variables--which is probably most important for the student to understand at this point, since he is sure to encounter the concept of correlation coefficients if he does very much reading in the social sciences.

One of the most useful ways to view a correlation coefficient--which measures the degree of relationship between two variables--is to see it as expressing the proportion of variation in one variable which is "explained" or accounted for by variation in another variable. It is this view of correlation which is stressed by Mueller and Schuessler and it is worth close attention.

From John H. Mueller and Karl F. Schuessler, Statistical Reasoning in Sociology (Boston: Houghton Mifflin Company, 1961) pp. 238-249; 281-310. Reprinted by permission.

John H. Mueller and Karl F. Schuessler,
Statistical Reasoning, pp 238-249; 281-310.

SECTION ONE

Concept of Statistical Association

Principle of Contingency. It is an axiom of science, as well as of common sense, that no event in nature "just happens," but always occurs under very specific, known or unknown, circumstances. An event is therefore never to be viewed in isolation. It must be considered as a product of the joint operation of many forces, each of which contributes a variable element to the observed outcome. Thus, the size of family may be dependent on such factors as age at marriage, level of income, extent of employment of the mother outside the home, and the religious ideology entertained by the parents. Parole success of released prisoners may be related to the type of crime committed, age, and history of recidivism. Some of these factors tend to promote and accelerate, others to retard in varying degrees or even inhibit, the occurrence of the event, or at least to modify its character or magnitude. Thus, the religious factor may foster large families, while the economic factor may restrain that tendency. In any case, we cannot predict the degree of parole success, or explain the size of the family, unconditionally, but only on the basis of specifically designated factors, or variables, on which the outcome is contingent. The human observer does not possess absolute knowledge; he must use one event as a cue to anticipate another. His understanding is therefore grounded in the *principle of contingency*. And it is in accordance with that principle that the sciences, both physical and social, have set up their methods and objectives: (1) to identify the variables (determining factors) associated with an event; (2) to uncover the patterns of association between the variables and the event; and (3) to measure the strength of that association.

Although no event is conceivable in isolation, it is not a simple matter to detect with what variables it is linked. For example, the association between delinquency and intelligence was once thought to be very close, but in later analysis was found to be quite remote. This shift in interpretation was a result not only of more carefully controlled studies of delinquency and its associated social factors, but also of a redefinition of the test-intelligence norms themselves.

To man, nature is infinitely complex; hence, the search for such underlying relations can never be completed. There are innumerable factors whose relations are so intricate that the "final" factor in the production of an event can never be attained, much less measured. However, in its unfulfilled quest for certainty, our common sense begins very early to construct patterns of relationships on the basis of which it endeavors to comprehend the past, understand the present, and thereby anticipate the future.

Patterns of Relationship. The patterns of association are, of course, continuously revised in the trial and error of daily experience. In the process, the observer mentally quantifies and summarizes his observations by first noting the factors which seem to him to "cause" or to be linked with the event and, second, by noting the frequency with which he successfully anticipates or forecasts it. In this casual manner, every person begins to cultivate the habits of association and becomes an informal statistician. He practices intuitively the principle of correlation. In fact, much sociology effectively employs such intuitive statistics, skillfully put together by alert, widely-traveled scholars, made without benefit of pencil and paper calculation or technical procedures. Indeed, many who are critical of the utility of statistical procedures nevertheless unwittingly employ them in this unofficial manner.

For some purposes, a rough subjective approximation of a correlation is fairly satisfactory, but for scientific purposes more accurate measurements are desired. Such precision is not a simple matter to achieve. The difficulty lies in the complexity of the patterns of relationship in terms of which we view and organize the world. Some of the salient features of this complexity in patterns may be formulated as follows: (1) every event is the outcome of multiple factors; (2) the force of these respective factors varies in intensity, and (3) may flow in one or more directions in producing its effects; (4) the factors are in constant interaction, and consequently (5) they may reinforce, counteract, and cancel one another. We illustrate with a popular example: the halfback in a football game, whose scoring power tends to be associated positively with running ability, is either obstructed or aided in actual performance by the condition of the field, his fatigue, the type of plays called, to say nothing of the varied activities of the other twenty-one players on the field. This association between scoring power and running ability may therefore not be visible to the fans in the stadium. Just so is the correlation between income and size of family beclouded by such factors as religion, age at marriage, occupation of breadwinner, and education. It may at first seem a hopeless task to disentangle these networks of relationships and to subject them to statistical reasoning. Nevertheless, in a large number of observations, the essential relation can be expected to shine through the haze formed by the multiplicity of factors.

This problem is probably less difficult in the physical than in the social sciences. The physical scientist, by means of available laboratory controls, is to a certain extent able to segregate and manipulate his elements and to replicate his careful and undisturbed observations, whereas the social scientist is often obliged to accept data which are like unrefined ores from "nature in the raw" and to assemble materials from widely dispersed sources and a variety of settings. He is therefore compelled to employ statistical controls, since, in general, laboratory controls are closed to him.

What Are the Evidences of Relationship? How can we be sure that factors are interconnected? And, having discovered a relationship, how may we determine the degree or intensity of that relation? Broadly speaking, there are two earmarks of such linkage: (1) joint occurrence of attributes, and (2) parallel changes in two or more series of quantitative observations.

The relative frequency with which certain attributes happen together is probably the most elementary basis of lay judgment of association. This is the *principle of joint occurrence*. Statistical variables, like human beings, are usually judged "by the company they keep." For example, if delinquency is more often found in boys than in girls, we conclude that delinquency is associated with "boyness." The strength of this association will vary according to other factors such as the boy's age, the type of delinquency, and many other elements in the pattern, all of which will render the statistical application of an apparently simple principle quite complicated. Hence, it need hardly be reiterated here that some system of tabulation and classification is necessary as an aid not only in establishing an association, but also in determining its strength.

Second, if in two series of quantitative data, a unit change in one variable is paralleled with some degree of regularity by a comparable change in the other series — that is, if they move together — we conclude that they are somehow tied together, and that there is an association between the two sets of data. For example, as income declines, the size of family tends to increase; and, if the observations endure through a rather extensive range — that is, for the entire range of families of all sizes and of incomes of varying amounts — the evidence of a relation is strengthened. This is called the *principle of covariation.*

John H. Mueller and Karl F. Schuessler

Devices for the Measurement of Association. Techniques for the measurement of association must be adapted (1) to the nature of the data, (2) to the number of variables being related, and (3) to the patterns of relationships among them. Hence, the devices will differ according to whether the data are in the form of attributes which are merely enumerated, or measures which are quantitatively scaled; and, additionally, whether the pattern of relationship between the variables is simple or complex.

Of the numerous devices available, we will present here six types, selected for (a) simplicity of calculation, (b) the prevalence of their use, and (c) their applicability to the data that are customarily encountered by sociologists. The different specimens illustrate a variety of approaches to the measurement of association and should above all convince the student of the relativity of their appropriateness.

The formulas, which are in some respects interrelated, may be conveniently grouped on the basis of the two principles of association already mentioned: the principle of joint occurrence and the principle of covariation. Under the first we include:

Yule's Coefficient of Association, Q
Phi Coefficient, ϕ
Coefficient of Contingency, C

Under the second, we include:

Rank-Difference Method of Correlation, ρ
Product-Moment Coefficient of Correlation, r
The Correlation Ratio, η

Questions and Problems

1. Define the following concepts:
    - Principle of Contingency
    - Pattern of Relationship
    - Joint Occurrence
    - Covariation
2. Characterize the following relationships as simple or complex and give your reasoning:
    - age at marriage and size of family
    - income and social status
    - infant-feeding and personality
    - capital punishment and incidence of murder
    - anomie and suicide
    - test and retest intelligence scores
    - diameter of circle and its circumference

SECTION TWO

Coefficient of Association: Yule's Q

One of the simplest measures of association is known as the ***Coefficient of Association***, or, more informally, as *Yule's Q*. This measure is designed to reflect the degree of association between a pair of qualitative variables, arranged in a 2 × 2 (fourfold) table. In his study of cross-classification the student has already acquired the sense of inferring association, and even cause-and-effect relationships, between dichotomous variables. But at that juncture, we did not seek to compute a single over-all measure which would reflect the strength of that relation. Yule's Q and other measures of association are designed to do just that, since they are summarizing measures for bivariate data, analogous to the mean for univariate data.

Inadequacy of the 2 × 1 Table for Estimate of Association. Many persons untrained in quantitative reasoning naively succumb to the temptation of drawing conclusions on the degree of association from a 2 × 1 table, instead of from a 2 × 2 table, which is the minimum for that purpose. The deceptive ease with which such erroneous deductions can be made is illustrated in the four 2 × 1 tables depicted in Table 9.2.1.

Table 9.2.1 *2 × 1 Tables*

Boys	*Per Cent*	*Republicans*	*Per Cent*
Delinquent	10%	Isolationist	50%
Non-delinquent	90	Internationalist	50
	100%		100%
Males	*Per Cent*	*Female Drivers*	*Per Cent*
A Grades	10%	Accident	20%
Other	90	No Accident	80
	100%		100%

Source: Hypothetical

Since only 10 boys out of 100 are delinquent, a layman may be misled into the conclusion that there is a very weak association between boyness and delinquency; similarly, that women are not accident prone. Since Republicans divide 50–50 on isolation, there would seem to be no striking tendency for Republicans to favor one or the other foreign policy.

But the important point is that delinquencies are distributed between boys and girls, and that auto accidents are shared by both men and women drivers. We must therefore inquire in what proportion the limited supply of violations or auto accidents are divided between the sexes. To which sex do delinquencies or accidents preferentially attach themselves? To estimate the degree of affinity for *either* sex, we must know the delinquency rate for both sexes; a delinquency rate for boys can be said to be high only when the rate for the girls is known. Let us therefore provide the rates for females and Democrats in the 2 × 1 tables, and analyze the results.

Table 9.2.2 *2 × 2 Tables*

	Boys	*Girls*		*Rep.*	*Dem.*
Delinquent	10%	0%	Isolationist	50%	30%
Non-delinquent	90	100	Internationalist	50	70
	100%	100%		100%	100%
	Male	*Female*		*Men*	*Women*
A Grades	10%	20%	Accident	10%	20%
Other	90	80	No Accident	90	80
	100%	100%		100%	100%

Source: Hypothetical

The delinquency rate of boys now turns out to be rather high, since even a 10 per cent delinquency rate is higher than no delinquency at all. Similarly, the grades of college men are below the norm; the Republicans display a marked propensity toward isolationism; and women drivers show a marked susceptibility to mishaps.

The fact is that we have now introduced a standard of judgment against which the 2 × 1 table may be compared. Some standard is inevitably and unwittingly introduced by every observer, and it is the function of statistical procedures to make the standard explicit. Hence, a 2 × 2 table is the minimum for a dependable conclusion on association.

The 2 × 2 Table. A 2 × 1 table merely presents the two subclasses of a single variable, whereas a 2 × 2 table presents the subclasses of two variables cross-classified by one another. In Table 9.2.3, a case is classified as boy and delinquent, as girl and non-delinquent. This double classification automatically establishes the distribution of joint occurrences between a given sex and behavior which is, of course, basic to the understanding of the phenomenon of association. In the hypo-

Table 9.2.3

Number of Delinquents by Sex

	Boy	Girl	*Total*
Delinquent	20	0	20
Non-delinquent	30	50	80
Total	50	50	100

thetical group of 100 children shown in Table 9.2.3, there are 20 delinquents, all of whom are boys. Most of the children are, of course, non-delinquent. But, given a delinquency, there is perfect prediction — that is, complete certainty about the sex of the delinquent. Hence, taking the data at face value, we may say that delinquency is completely explained by boyness, since there is obviously no element of girlness required for its occurrence. The completeness of this explanation should register itself in any index that may be contrived for that purpose.

In order to provide a single measure of association for such a 2 × 2 table, the English statistician, G. Udney Yule, proposed the following coefficient of association, which he labeled "Q" in honor of the nineteenth century statistician, Quételet:

$$Q = \frac{ad - bc}{ad + bc} \tag{9.2.1}$$

where a, b, c, d are the joint frequencies conventionally arranged in the fourfold table in the following manner:

a	b
c	d

Applying the formula to Table 9.2.3:

$$\begin{aligned} Q &= \frac{(20 \times 50) - (30 \times 0)}{(20 \times 50) + (30 \times 0)} \\ &= \frac{1000 - 0}{1000 + 0} \\ &= 1 \end{aligned}$$

This index of unity — a result already anticipated — is the obvious measure of the *complete* association between maleness and delinquency, since all delinquents are male, and all females are necessarily non-delinquent.

Sensitivity of Q. Any index must be able to discriminate to some extent between slight variations in the data. It will later become apparent that each of the various measures of association has its own idiosyncrasies in reflecting gradations of association. Yule's Q, being our first example of a measure of association, in its simplicity supplies us with a convenient introduction to this type of analysis which will, in its turn, throw considerable light on all subsequent instances. How sensitive, then, is Q to the variations in the data?

First let us assume that not all the delinquents are boys, but rather that 1 and 5 girls in 50, respectively, are delinquents. (See *a* and *b* of Table 9.2.4.) In both instances, there is still a positive association between delinquency and boyness, but not as complete as in the first table. Thus, the value of the index is reduced; it reaches zero when a delinquent is just as likely to be a boy as a girl. Sex and delinquency are then said to be independent; pure chance prevails. Statistically speaking, the internal cell ratios are identical with the corresponding marginal ratios. You might as well toss a coin for your prediction; hence, $Q = 0$.

Table 9.2.4 *Sensitivity of Q*

	(*a*)			(*b*)			(*c*)		
	B	G		B	G		B	G	
Del.	19	1	20	15	5	20	10	10	20
Non-del.	31	49	80	35	45	80	40	40	80
	50	50	100	50	50	100	50	50	100
	$Q = .94$			$Q = .59$			$Q = 0$		

Effect of Marginal Ratios. We have seen how Q reflects the changes in row (column) ratios, with marginals remaining constant. However, Q is not at all sensitive to changes in marginal ratios, so long as the corresponding cell ratio remains fixed. For purposes in hand, let us construe the column subtotals as samples of 100 boys and 50 girls (Table 9.2.5).

Table 9.2.5 *Stability of Q*

	(*a*)			(*b*)			(*c*)		
	B	G		B	G		B	G	
Del.	10	3	13	10	6	16	20	6	26
Non-del.	90	47	137	90	94	184	90	47	137
	100	50	150	100	100	200	110	53	163
	$Q = .27$			$Q = .27$			$Q = .27$		

One might intuitively suppose that the observed association of $Q = .27$ may be due in part to the fact that there are twice as many boys as girls in the total sample of 150, and that by doubling the sample of girls the index might increase. Let us, therefore, double the number of girls, retaining, however, the same *rate* of delinquency among the girls. It will be observed that the value of Q remains unchanged under this inflation. Similarly, if we double the supply of delinquents from 13 to 26 by multiplying the frequencies of that row by 2, the value of Q again remains unchanged.

In general, therefore, a change in the relative size of the marginals does not affect the value of Q, so long as the ratios within either columns or rows remain undisturbed. As we shall see later, this does not hold for ϕ, another index of association for fourfold tables. In this respect Q is more stable than is ϕ. Whether this type of insensitivity is desirable — and should therefore be considered a virtue of stability — or whether it is a defect of the formula, must be left to the good judgment of the worker in terms of his purposes. It is doubtless important to know that the number of observations in a subclass or stratum does not, of itself, influence the size of Q.

Definition of Association: Observed and Chance Frequencies. The above illustrative tables yielded, among other things, the important generalization that, when the internal cell ratios and corresponding marginal ratios coincide, and there is no association, the index must be zero. Only when these two sets of ratios differ, and to the extent of that difference, is there statistical association. Indeed, the formula for Q rests on this principle of the discrepancy between the observed and chance cell frequencies.

The chance frequencies are easily derivable from the marginal frequencies according to the following logic: the boys constitute 50 per cent of the children; hence, if the two sexes are equally susceptible to delinquency (or, in other words, if the sex factor has no influence on the production of delinquency — i.e., if sex and delinquency are independent), the boys would also have 50 per cent of the delinquencies. To express this in still another way, if delinquencies were divided impartially (by chance) among the two sexes, the boys' quota would be 50 per cent, or in terms of our example, the boys would have 10 of the 20 delinquencies. Arithmetically put:

$$\frac{50}{100} \times 20 = 10$$

In essence, we simply adjust the cell ratios to be identical with the corresponding marginal ratios (Table 9.2.6). The boys actually show 5 more delinquencies than would be expected by chance, while the girls show

Table 9.2.6 *Observed and Chance Frequencies*

	Observed			Expected		
	B	G		B	G	
Del.	15	5	20	10	10	20
Non-del.	35	45	80	40	40	80
	50	50	100	50	50	100
	$Q = .59$			$Q = 0$		

the corresponding deficiency. There is therefore a positive affinity between boys and delinquency. However, since the excess of 15 over 10 is fairly moderate, the index reveals only a moderate association: $Q = .59$. But it is the amount of this excess that constitutes, by definition, the degree of association. This difference between the observed and expected frequencies is conventionally symbolized as $(O - E)$ — an expression that has achieved an almost epigrammatic currency in statistical methods of investigation, which the student will learn to appreciate.

Signs. Although mathematically the formula for Q will yield a sign, it may not seem very meaningful when applied to attributes such as sex, religion, or race, which do not constitute a hierarchy. In the foregoing illustration, therefore, one may say that maleness is positively associated with delinquency, but it cannot be verbalized as "the higher the sex, the higher the delinquency rate." Nevertheless, we may be guided by the algebraic sign which always refers to the association indicated in cells a and d. When the sign is positive, the joint frequencies of the attributes intersecting in cells a and d exceed chance (and are therefore linked together); when the sign is negative, the frequencies in cells a and d are less than would be expected by chance, and therefore "repel" each other. By corollary, cells b and c are read in the opposite manner. The assumption is, of course, that the alphabetical designations of cells are made in the conventional order.

Type of Data to Which Q Is Applicable. While Q is singularly appropriate for genuine attributes, nevertheless this measure may at times be applied to continuous variables by prudently cutting them into dichotomies. For example, persons may be subclassified as under and over 21 years, and responses may be either positive or negative, although these latter usually shade into each other (see Table 9.2.7).

There are, however, two cautions which should always be observed

Table 9.2.7

Opinion by Age

		Opinion Yes %	*Opinion* No %
Age	−21	60	40
	21+	40	60
		100	100

in such dichotomization. (1) Such compression of a mass of continuous, detailed observations into two broad dichotomies may be an expensive waste of perfectly good data, and naturally reduces the potential precision of the resulting measure. Instead of cavalierly discarding costly accuracy, it might be advisable to use other available measures which can take such precision into account. (2) A less obvious trap for the unwary operator is the more or less arbitrary location of the cutting point, which introduces an unpredictable effect upon the ultimate index. If, in the above table, the cutting point had been set at 30 years instead of 21, the cell frequencies might have been radically changed, and the Q-measure substantially altered. Such arbitrary decisions gravely reduce the reliability and comparability of any measure, including Q. The student will remember the general principle that all grouping has its hazards; but when everything is staked on one cutting point, the precariousness of the undertaking is increased.

The usefulness of Q also varies with the general pattern of distribution of the data. These patterns may be particularized as follows: (1) the division of frequencies within each marginal set; (2) the similarity (dissimilarity) between the two marginal sets; and (3) the internal distribution of cell frequencies. Thus, an individual marginal set may be symmetrical (equal division) or skewed; the two marginal sets may be more or less identical in degree of symmetry; and the cell frequencies may be approximately evenly dispersed in all four cells, concentrated in three cells, or in two diagonal cells (see Tables 9.2.8a and 9.2.8b).

It will become increasingly evident that Q is most useful, and least

Table 9.2.8a *Like Marginal Sets*

30	25	55	5	20	25	0	50	50
25	25	50	70	5	75	50	0	50
55	50	105	75	25	100	50	50	100

Table 9.2.8b *Unlike Marginal Sets*

50	0	50		5	40	45
25	25	50		15	40	55
75	25	100		20	80	100

competitive with other measures of association, in those situations where the cases fall predominantly in three of the four cells. Such a distribution, which is evidence of a one-way association, is possible only when marginal sets are unlike. Hence, Q is a likely choice whenever marginal sets are dissimilar. The reason for this will become clearer after ϕ and other indexes of association have been studied.

Interpretation of Yule's Q. For all its apparent simplicity and precision, Q possesses no simple quantitative meaning, and is not convertible into a specific prediction.* For example, if the association between sex and delinquency is .5, one cannot say that 50 per cent of a given sex is delinquent, or that one sex is twice as delinquent as another, or that sex accounts for 50 per cent of all delinquency. A Q-value of .8 is not twice as strong as .4, although it is stronger than all values less than .8.

Not even is a coefficient of 1.00 wholly unambiguous. In the previous example, the perfect Q did not depend on the tendency of boys to become delinquent, but rather on the relative tendency that a delinquency be committed by a boy. Actually, in absolute terms, the tendency of a boy to be delinquent could be very weak, and Q still be 1.00, provided that no girls are delinquent. A coefficient of 1.00 therefore registers the presence of one-way association, but does not reflect the degree of prediction in the other direction, although that too may be complete. We may say, therefore, that Q is primarily a measure of one-way association.

In general, it would seem to be sound practice to quote the original 2 × 2 tabulation along with the calculated value of Q. Not only is the table necessary in order to identify linkages between specific attributes; it also permits the reader to examine for himself the detailed structure of the relationship. But if it is a single index that is required, then one must accept the limitations of such a condensation, just as the mean was accepted as a limited representative of the full array.

---

* A precise probabilistic interpretation of Q has recently been formulated. See L. A. Goodman and W. H. Kruskal, "Measures of Association for Cross Classifications," *Journal of the American Statistical Association*, XLIX, 1954, p. 750.

John H. Mueller and Karl F. Schuessler

SECTION TWO

Scatter Diagram and Correlation Table

In the previous section, we were concerned with the problem of measuring the amount of agreement between ranks, or ordinal measures. However, there are numerous concepts in sociology that are subject to interval measurement: for example, we may measure fertility in terms of the number of births per 100 women or wages in terms of dollars and cents. When the data being correlated consist of scaled variables, we naturally employ techniques that are appropriate to that type of quantitative data. While these correlational techniques differ somewhat from those previously presented, they all answer to the same purpose: namely, to express as precisely as possible the degree of relationship between two or more variables. The ensuing discussion is restricted to the most prevalent of all correlation measures: the product-moment correlation coefficient (r) and the correlation ratio (η).

The Scatter Diagram. The measurement of covariation can be approached in a preliminary manner by means of the *scatter diagram*, which is related to *bivariate* data in much the same way as the histogram is to univariate data. It reveals at a glance the entire disposition of items, thereby enabling us to arrive at a rough but useful estimate of the strength of the correlation before we actually measure it. This indispensable device has already been employed to depict the trend of a time series (p. 99) and the degree of correspondence between ranks (p. 274); however, the details of its construction have not yet been elaborated.

To illustrate more fully its construction and use, we take as our point of departure Table 10.2.1, which presents yearly income averages and suicide rates for the U.S. Central states, 1929–1950. To reduce the data of this table to a scatter diagram, we first draw horizontal and vertical axes, as in the construction of a histogram. Axes are drawn approximately equal in length, unless there is good reason to deviate from this convention. Also, as a matter of convention, the independent X-variable is plotted along the base line, and the dependent Y-variable along the vertical axis. The establishment of one variable as the independent, and the other as dependent, is, of course, not a statistical problem, but rather a matter of judgment and circumstance. The dependent variable may be construed as the effect of the independent variable as cause, or the outcome to be predicted from the predictor variable. In many instances, there may be no clear "causal" dependency at all, since both variables may be viewed as the consequence of an unidentified third factor.

Scales are next established on the axes in such a manner as to accom-

Table 10.2.1

Suicide Rate by Annual Per Capita Income, Central U.S., 1929–1950

Year	Per Capita Income (X)	Suicide Rate (Y)
1929	604	15.4
1930	532	17.9
1931	486	19.1
1932	399	19.3
1933	400	18.1
1934	437	16.9
1935	482	15.5
1936	549	15.2
1937	567	16.3
1938	515	16.7
1939	566	15.3
1940	602	15.2
1941	708	14.0
1942	800	12.8
1943	914	11.0
1944	972	11.0
1945	985	12.3
1946	951	12.3
1947	897	12.1
1948	928	11.6
1949	878	11.9
1950	947	11.9

Source: Donald Faigle, *Suicide in Relation to Income, Urbanization and Race*, Unpublished Master's Thesis, Department of Sociology, Indiana University, 1956 Table 5.

modate, with a margin to spare, the observed ranges of the respective variables. Thus, the horizontal scale covers the distance from $200 to $1,200, while the vertical scale extends from 10 to 20. Needless to say, enough markers are set up on each axis to ensure accurate and effortless plotting. Unlike the histogram, the vertical scale in a bivariate plot need not begin with zero, for the reason that the focus of attention is on the contour of the scatter rather than on relative frequency as gauged by the height of the curve. In this respect, the scatter diagram is analogous to the semi-log time chart, whose meaning likewise inheres in its shape rather than location.

Having drawn and scaled the axes, we are ready to plot each pair of values by a double-duty point. For any pair of variates, the Y-value fixes the height of the point above the base line and the X-value fixes its horizontal distance from the vertical axis. Thus, 1940 is represented in Figure 10.2.1 by a point located at the intersection of guide lines ex-

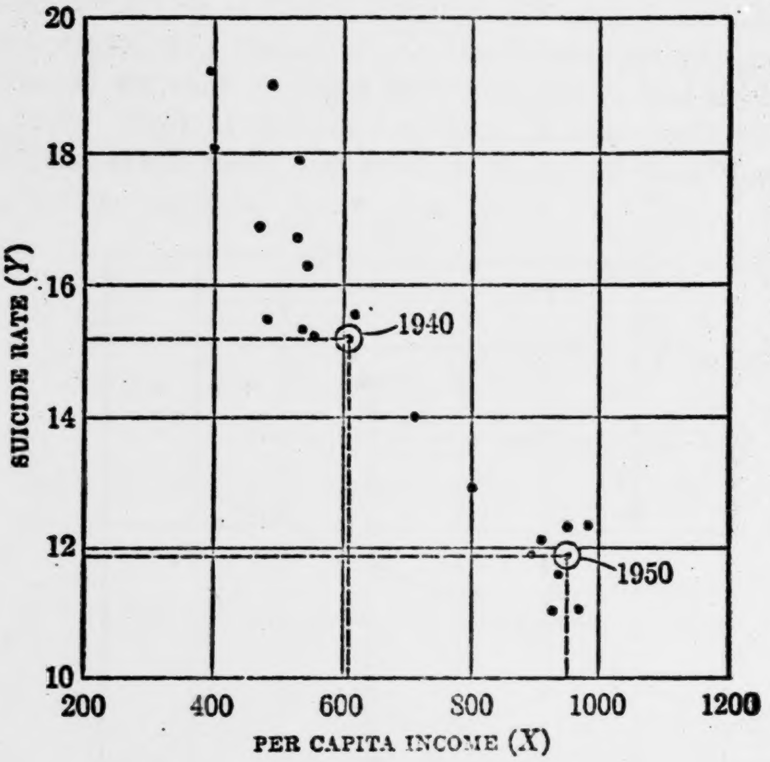


FIGURE 10.2.1 *Scatter Diagram, Suicide Rate by Annual Per Capita Income, Central U.S., 1929–1950*

tending perpendicularly from $Y = 15.2$ and $X = 602$; similarly, 1950 is represented by a point at the intersection of 11.9 and 947. The swarm of all such points constitutes the scatter diagram.

Types of Scatter. It is the pattern of this swarm that enables us to judge the nature of the relationship — and such a judgment is an essential preliminary to its proper measurement. Thus, it appears that a fixed increase in income is accompanied by a fixed decrease in the suicide rate; that is, the suicide rate changes by a constant amount per unit income. Such a relation is termed *rectilinear*, or simply *linear*, because the trend of scatter conforms to the track of a straight line.

Any such trend line, whether freehand or mathematically fitted, is technically termed a *line of regression*. This concept was coined by Galton in 1877, who used it in connection with his correlational studies of the characteristics of parents and their offspring. He perceived that such a line effectively expressed the tendency of children to "regress" to the average level of the parents in a wide variety of traits. The term has survived and enjoys a wide currency, although it is no longer restricted to its original connotation.

Furthermore, the relation in Figure 10.2.1 is said to be ***inverse***, since the two series move in opposite directions: as income rises, the suicide rate falls. If suicide and income had risen together, then the trend of points would have been upward, reading from left to right. Such a sloping trend line (Figure 10.2.2) is evidence of a *direct* linear relationship between two variables; family size increases as farm acreage increases.

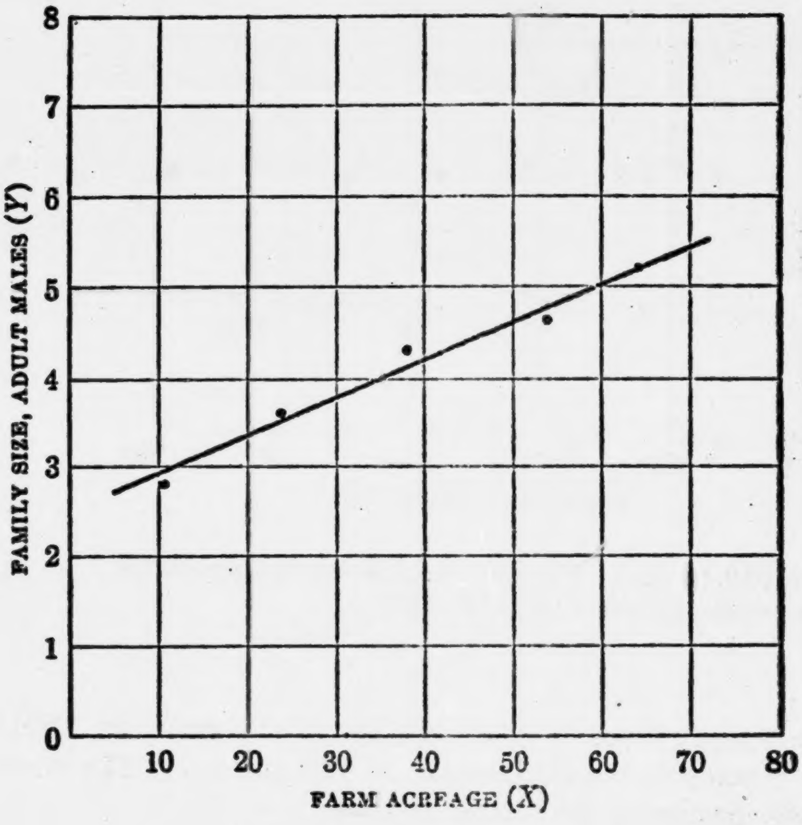


Figure 10.2.2 *Family Size by Farm Acreage*

The trend of the scatter will not always be linear; rather it may be *curvilinear* and take on any one of innumerable curve patterns. A simple example is provided in Figure 10.2.3, which portrays the relation between favorableness of attitude toward a minority people and intensity of feeling. As might be anticipated, a decided opinion — whether pro or con — is held with considerable intensity of feeling, while a less decided or neutral opinion arouses no very strong feelings. Whether this pattern — strong opinion, strong affect; weak opinion, weak affect — holds under all conditions is not here our concern. That would be a matter for empirical investigation. Here we are merely interested in exhibiting a type of relationship which is fairly common in sociological studies.

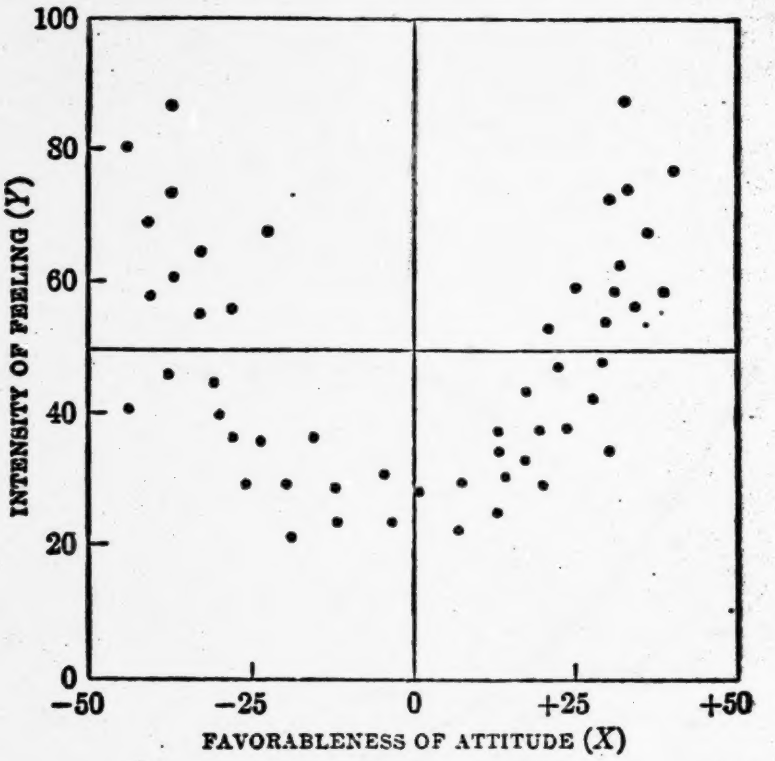


Figure 10.2.3 *Scatter Diagram, Intensity of Feeling by Favorableness of Attitude* (*Hypothetical Data*)

The three scatter diagrams presented thus far have distinctive trends and one would have little hesitancy in describing them as exclusively linear or curvilinear. But scatters of empirical observations are seldom so clean and unambiguous; more often both linear and curvilinear tendencies combine in the same set of data and thereby complicate the problem of representing correlation by an over-all measure. For example, the scatter of delinquency rates and average monthly rentals for 140 small census tracts in the city of Chicago (Figure 10.2.4) is marked by some linearity, yet it appears that a curved trend line would better fit the entire scatter. From left to right, as rentals increase, the delinquency rate responds by decreasing, but at a progressively slower rate. This is evidenced by the straightening of the swarm. There are of course several striking exceptions to the foregoing generalization — a few extremely high delinquency rates occur with above-average rentals. These mavericks would require special analysis, since they represent a breakdown in the "law." Yet, in the main, the law of relationship holds fairly well, affording a measure of predictability of one variable from the other. If we knew, for example, the rental to be $60, we would predict the delinquency rate to be approximately 3.0, which is the height of the free-hand *regression curve* at that point. To be sure, such a prediction would not

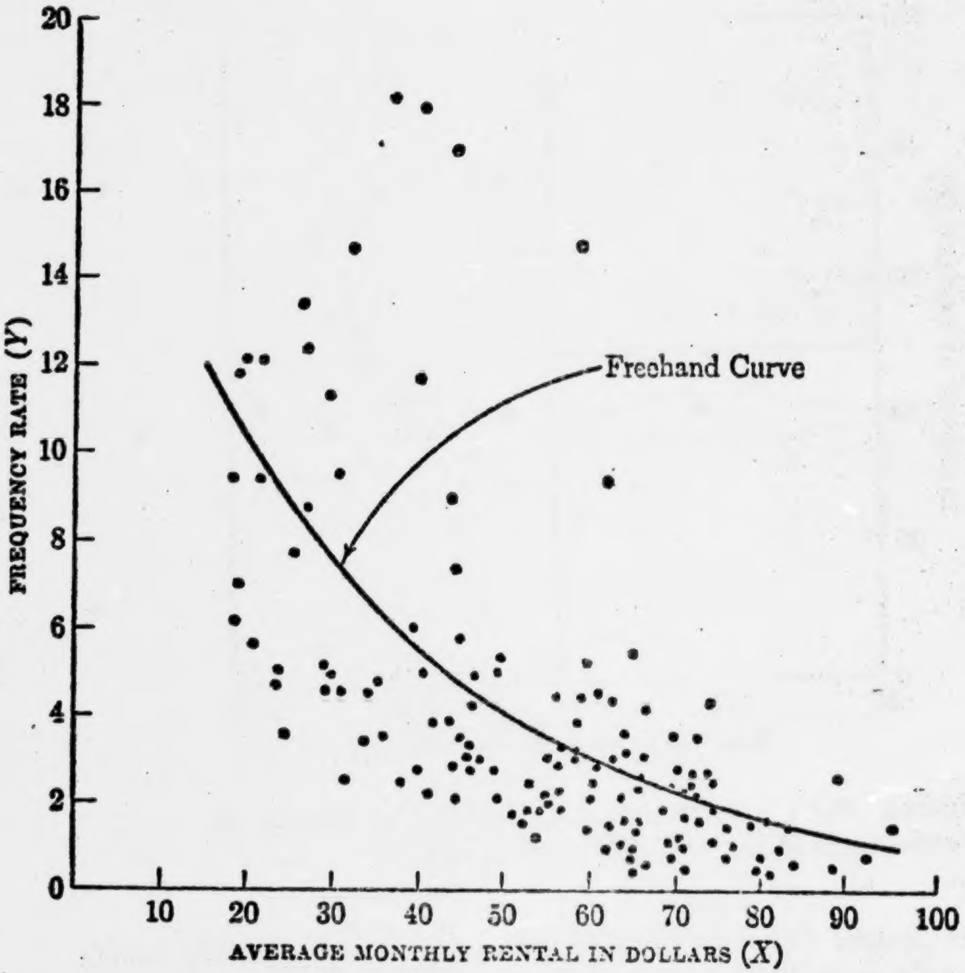
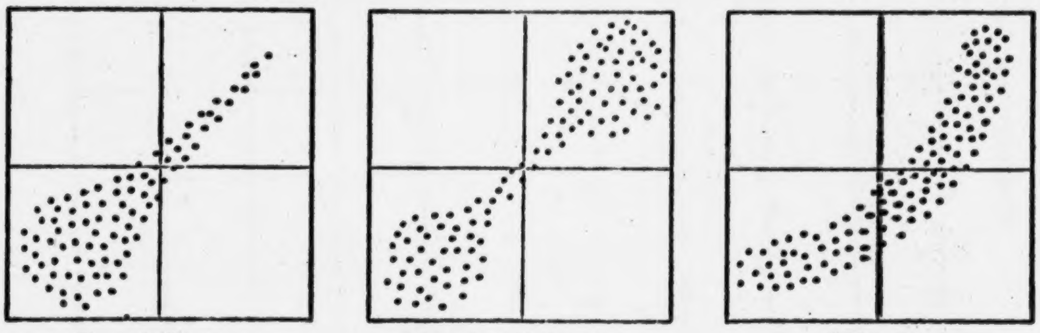


Figure 10.2.4 *Scatter Diagram and Freehand Regression Line, Delinquency Rate by Monthly Rental, 140 Local Areas, Chicago, 1930*

be free of error, for the obvious reason that none of the observed values fall right on the curve at that point — all deviate to a greater or lesser extent. Evidently, the accuracy of any such prediction would vary according to the tendency of the points to hug the line of relationship between the two series. When the points move within a narrow lane, predictive accuracy, and therefore correlation, would be high; when the points are widely scattered, predictive accuracy would be correspondingly low. Only when all points fall right on the regression line would prediction and correlation be perfect. At the other extreme, when the scatter is purely random, then we may just as well ignore the so-called "predictor variable." For any or all rentals, our best guess would be the over-all mean of the delinquency rates.

Scedasticity. Knowing the rental to be $30, we still could not accurately forecast the level of delinquency, since the delinquency values are widely scattered for that value; on the other hand, for a rental of $70, we could predict with much greater accuracy, owing to the bunching of delinquency rates around the regression line. This scatter of Y-values for the respective X-values is known as *scedasticity*. If the degree of variation in delinquency rates — the width of the scatter band — had been uniform for all values of X, then we could have spoken of Y as being *homoscedastic* in respect to X. Actually, the degree of scatter in Y diminishes as X changes, so that Y is *heteroscedastic* in respect to X. Heteroscedasticity implies that the degree of correlation is not uniform throughout the entire series; hence, its presence reduces the feasibility of a single over-all measure of correlation, which is after all an average. Just as we hesitate to compute the mean of heterogeneous bimodal data, similarly we hesitate to calculate an average measure of correlation of a heteroscedastic scatter.

Heteroscedasticity may be even more glaring than that exhibited in Figure 10.2.4; the scatter may be gourd-shaped, dumbbell-shaped, or J-shaped, as in Figure 10.2.5. These oddly shaped scatter diagrams by no means exhaust the variety of types that may be encountered in practical work. However, they do serve to ratify the utility of this visual aid. Although the scatter diagram yields no mathematical measure of correlation, it does indicate (a) whether the relationship is simply rectilinear or more complex, (b) whether or not the relationship is consistent over the entire range, and (c) whether the relation is strong or weak. It is an indispensable tool and plays the same role in correlation as does the frequency graph in the processing of univariate data. It provides a bird's-eye view of the whole distribution.



Gourd-Scatter Dumbbell-Scatter J-Scatter

FIGURE 10.2.5 *Selected Types of Scatter*

The Joint Frequency Table. Instead of plotting the individual items of the bivariate data, we may group them. Such grouping answers the general purposes of all grouping: (1) to reveal the basic pattern of distribution, and (2) to facilitate the statistical processing of the data. In grouping bivariate data, we classify each case simultaneously in two class intervals, thereby locating each case at the intersection of a given row and given column. Hence, it is as though we superimposed a grid on the scatter diagram, counted the points in each cell and inserted the corresponding number. Such an operation, applied to the scatter of delinquency rates and average rentals (Figure 10.2.6), would yield the joint frequencies of Table 10.2.2.

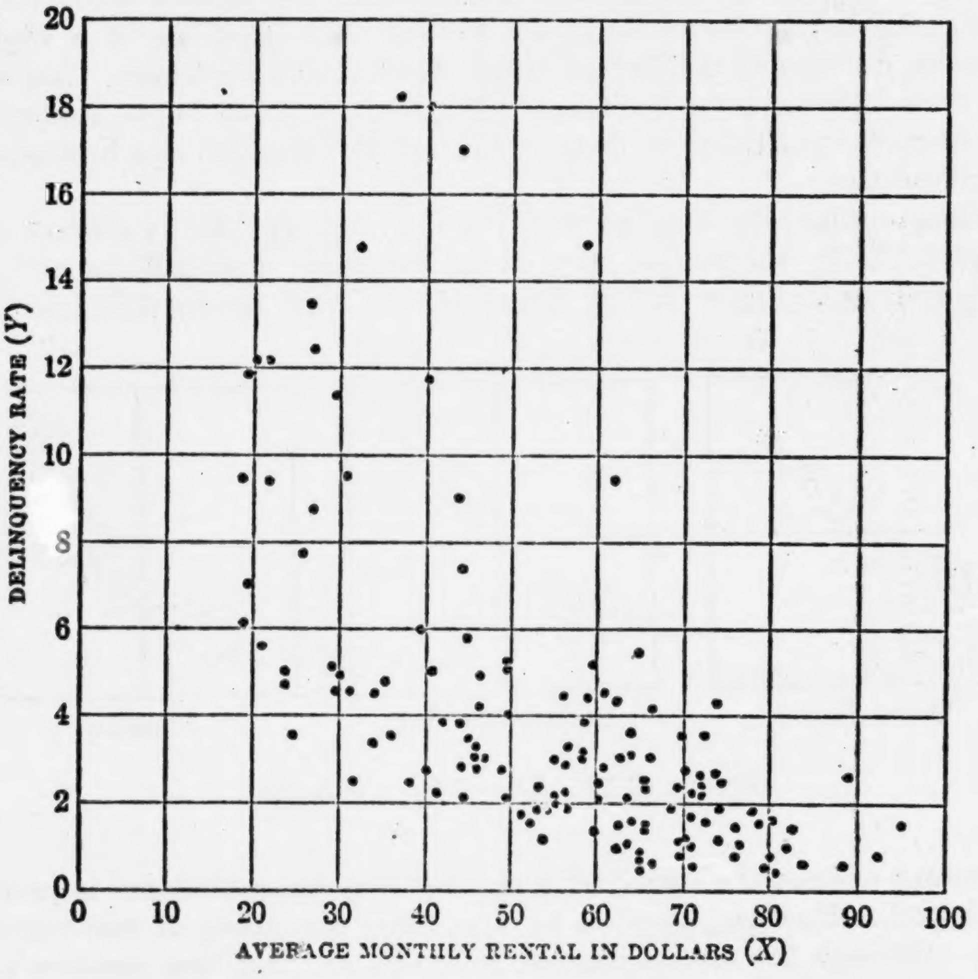


Figure 10.2.6 *Grid on Scatter Diagram, Delinquency Rate by Average Monthly Rental*

Naturally, in any real situation we would not proceed in this somewhat fanciful manner; rather, we would immediately tally the unarrayed bivariate items in the grid set up for that purpose, and then count the joint occurrences in each cell.

Table 10.2.2 *Joint Frequency Table; Delinquency Rates by Average Monthly Rentals*

DELINQUENCY RATE ($\bar{Y}$)	AVERAGE MONTHLY RENTAL IN DOLLARS ($\bar{X}$)										Frequency (f_y)
	0–9	10–19	20–29	30–39	40–49	50–59	60–69	70–79	80–89	90–99	
18.0–19.9				1	1						2
16.0–17.9					1						1
14.0–15.9				1		1					2
12.0–13.9			4								4
10.0–11.9		1	1		1						3
8.0– 9.9		1	2	1	1		1				6
6.0– 7.9		2	1	1	1						5
4.0– 5.9			5	4	6	3	4	1			23
2.0– 3.9			1	4	14	11	12	9	1		52
0– 1.9						7	13	13	7	2	42
Frequency (f_x)	0	4	14	12	25	22	30	23	8	2	140

Grouping Procedure. To clarify this grouping procedure, we list the directives observed in the construction of Table 10.2.2.

1. Select a suitable class interval for each variable.
2. Mark off these class intervals on properly drawn axes, and from these markers extend guide lines in such fashion as to create a gridwork of cells. Each cell is taken to represent the junction of two class intervals.
3. Locate each pair of values in its proper cell and indicate its presence there by a tally mark.
4. Count tally marks in each cell and replace tally by that number. Any such number represents a joint frequency—i.e., the frequency with which two class values or midpoints, occur together.
5. Sum by rows and columns in order to obtain the marginal totals. These marginal totals of course constitute the simple frequency distributions of the individual variables.
6. Sum marginal frequencies to obtain the grand total of cases, N, the one sum serving as a check on the other.

In retrospect, we see that the construction of the joint frequency table proceeds according to well-established principles: the classification should be fine enough to reveal the shape of the joint frequency distribution, but not so fine that some rows and columns are completely vacant. But if one is to err at the outset, it is perhaps preferable to have too many cells than too few. The optimum arrangement will be more apparent in the light of too much detail rather than too little, and it is usually easier to amalgamate cell frequencies than to partition them.

Function of the Joint Frequency Table. We must acknowledge that a joint frequency table cannot depict the type of relationship so vividly as the *scattergram,* for the reason that numerals cannot convey so effectively gradations of density as does the scatter of dots. Nevertheless, in spite of its relative coarseness, it will often be more than adequate for purposes of exhibiting the pattern of relationship and assuring its comprehension.

Moreover, the column and row distributions of the joint frequency table permit a statistical measurement of scedasticity which would be impossible to obtain from a raw scatter diagram. For such an assessment, we need only compute the standard deviation of each column or row. A close similarity among such standard deviations would be evidence of homoscedasticity, whereas pronounced differences would suggest heteroscedasticity.

Furthermore, the marginal distributions of the *X*- and *Y*-variables — which, of course, are integral features of the joint frequency table — also supply important and recognizable clues to the possible degree of association between the plotted variables. Specifically, the marginal distributions set limits to the degree of obtainable correlation. For example, unlike marginals preclude a perfect linear relationship. This merely generalizes the earlier observation in respect to the ϕ-coefficient that perfect two-way association in contingency tables is impossible so long as marginal sets are not identical. This general principle is illustrated by Table 10.2.3, whose marginals make it impossible to locate

Table 10.2.3

Unlike Marginal Sets

				f_y
Y-Variable				10
				10
				10
f_x	15	10	5	30
	X-Variable			

all cases along one diagonal, as would be required in perfect rectilinear correlation. No matter how we deploy the cases, we cannot force them onto one diagonal, as the student should experimentally demonstrate for himself. In general, the marginals always act to constrain in one way or another the type and degree of association between paired variables; hence, it is always necessary to examine them in order to determine the limits which they impose, and the prospects for the valid employment of a given correlation index.

Additionally, the joint frequency table may serve as a computing aid which may be called on especially when the number of instances is large or when computing machines are not available. Finally, it is a prerequisite to the simplest approach to curvilinear correlation, as will be evident in a later section.

This section has stressed the important rôle of the scatter diagram and the joint frequency table in the preliminary analysis of covariation. From these charts it is possible to determine whether the relationship is (a) linear or curvilinear; (b) direct or inverse; (c) weak or strong; additionally, (d) whether the variables are homoscedastic in respect to each other; (e) whether there are notable exceptions to the main trend or "law of relation," and (f) whether the marginal distributions are symmetrical or skewed, and approximately matched. Because of these significant disclosures, it is mandatory that, prior to the measurement of correlation, a scatter diagram or joint frequency table, or even both, be constructed and carefully studied. Without these visual aids, we fall into the danger of using improper procedures and thereby arriving at misleading conclusions.

QUESTIONS AND PROBLEMS

1. Define the following concepts:
    - Scatter Diagram
    - Bivariate Data
    - Rectilinearity
    - Scedasticity
    - Curvilinearity
    - Joint Frequency Table
    - Marginal Distributions
2. If the marginal distributions are normal in shape, will the scatter necessarily be homoscedastic? Illustrate and explain.
3. Is a perfect rectilinear scatter possible, if marginal distributions are unlike? Illustrate your answer by table or graph.
4. Does the presence of homoscedasticity in a scatter guarantee linearity? Illustrate by sketch.
5. Does perfect rectilinearity in a scatter require identical marginals? Illustrate.

Table 10.2.4

Clothing Expenditures by Total Family Expenditures, U.S.

TOTAL EXPENDITURES ('00 OF DOLLARS) (X)	CLOTHING EXPENDITURES (PER CENT OF TOTAL) (Y)
9	**5**
10	**8**
14	**12**
18	**13**
20	**14**
24	**15**
28	**16**
34	**18**
43	**20**

Table 10.2.5

Per Cent Democratic Vote by Voter Registration, Selected States, Congressional Elections, 1952

STATE	DEMOCRATIC VOTE (% OF TOTAL)	PER CENT REGISTERED
Ariz.	52	78
Ark.	85	50
Cal.	44	88
Colo.	44	91
Conn.	46	91
Fla.	74	75
Ga.	100	59
Idaho	41	96
Ind.	43	96
La.	91	67
Md.	48	62
Mass.	46	88
Mont.	43	83
Nev.	50	81
N.H.	37	92
N.J.	42	85
N.Y.	40	80
Ohio	44	59
Ore.	39	87
Pa.	48	78
R.I.	54	87
S.C.	98	49
S.D.	31	83
Vt.	28	88
Va.	67	33
Wash.	43	92

Source: U.S. Bureau of the Census, *Statistical Abstract of the U.S., 1952*, U.S. Government Printing Office, Washington, D.C., 1952

6. If a scatter is homoscedastic, does it follow that the relationship is close?

7. (a) Construct a scatter diagram for the data in Table 10.2.4.
   (b) Formulate in your own words the "law of relation."

8. (a) Plot a scatter diagram for the following data of Table 10.2.5.
   (b) Characterize the relationship. Does a preponderant one-party vote appear to depress voter registration?

9. Construct the scatter diagram for the delinquency and recidivism rates (Table 10.2.6). How would this trend be described in words?

Table 10.2.6

Delinquency and Recidivism Rates, Chicago, 1929

Delinquency Rate per 100	Recidivism Rate
0.5	26.1
1.5	26.6
2.5	35.1
3.5	30.1
4.5	38.7
5.5	54.8
6.5	42.9
7.5	54.7
8.5	32.4
9.5	52.3
10.5	66.7
11.5	50.2
12.5	59.9
13.5	65.9
15.5	53.6
16.5	80.3
19.5	61.1

Source: Clifford R. Shaw, *Delinquency Areas*, The University of Chicago Press, 1929, Table 21, p. 181.

10. (a) Plot a scatter diagram from Table 10.2.7.
    (b) What does the scatter of this graph suggest concerning the "pull" of economic conditions? Explain the two fairly distinct clusters.

11. Construct a joint frequency table for the data in Table 10.2.7. Divide Income Index into ten class intervals of width 10, lower rounded limit 50; divide Per Cent Born in Other States into ten class intervals of width 10, lower limit 0. Tally marginal frequencies.

12. Sketch the marginal distributions which would yield the scatters shown in Figure 10.2.7.

John H. Mueller and Karl F. Schuessler

Table 10.2.7 *Per Cent Born in Other States by Income Index, U.S.*

STATE	INDEX OF INCOME *	% BORN IN OTHER STATES	STATE	INDEX OF INCOME	% BORN IN OTHER STATES
Ala.	62	11	Neb.	96	25
Ariz.	91	44	Nev.	137	68
Ark.	58	22	N.H.	99	30
Cal.	124	58	N.J.	120	32
Col.	99	50	N. Mex.	81	44
Conn.	127	28	N.Y.	124	17
Del.	138	35	N.C.	64	12
Fla.	80	54	N. Dak.	75	23
Ga.	69	14	Ohio	115	24
Idaho	88	49	Okla.	78	39
Ill.	121	25	Ore.	106	57
Ind.	103	25	Penn.	104	12
Iowa	94	19	R.I.	101	24
Kansas	104	32	S.C.	67	11
Ky.	69	12	S. Dak.	77	29
La.	74	15	Tenn.	69	20
Me.	83	12	Texas	89	21
Md.	107	34	Utah	88	21
Mass.	107	15	Vt.	82	22
Mich.	111	28	Va.	81	25
Minn.	91	21	Wash.	110	54
Miss.	50	11	W. Va.	75	17
Mo.	97	25	Wis.	101	15
Mont.	104	44	Wyoming	98	60

* Ratio of state per capita income to national per capita income.

Source: U.S. Bureau of the Census, *Statistical Abstract of the U.S., 1954*, U.S. Government Printing Office, Washington, D.C., 1954, Tables 37 and 337.

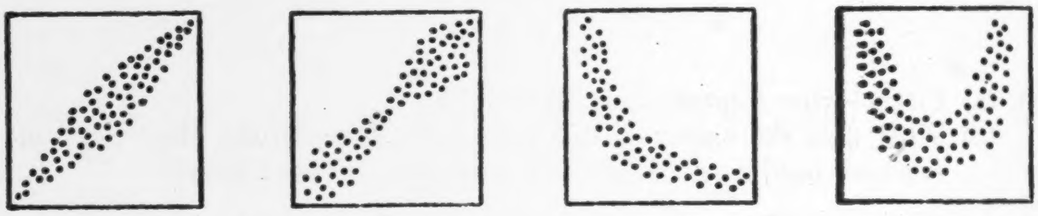


FIGURE 10.2.7 *Selected Scatter Patterns*

SECTION THREE

Linear Correlation of Two Variables

The Need for an Over-All Measure of Correlation. The scatter diagram, which was elaborated in the preceding section, reveals to the eye whether two variables change together in a systematic manner. The cluster of plotted points around the hypothetical regression line suggests the "law of relation" between the two variables. Moreover, by observing the width of the scatter, we can make at least a preliminary judgment as to how well the cases conform to this hypothetical law. Such conclusions are often of considerable value; yet they still leave something to be desired, for they are impressionistic, subjective, and unstandardized. Consequently, they cannot be described or communicated to anyone else without reproducing the chart — or at least engaging in an extended account of it. This is, of course, impractical. And since these "eye-measures" are not accurate in a mathematical sense, comparisons with similar measures are impossible even if all the intricate charts are available. What we need therefore is an objective, standard, synoptic *measure* of the relation between the two variables. This will finally be the measure of correlation.

In principle, there is of course no difference between rough measures made by the eye and those computed by a mathematical operation. The latter are merely more precise and serviceable. What we see is a hypothetical trend line to which the swarm of points more or less conforms and from which we deduce the degree of correlation. But in order to measure this correlation we must (1) establish exactly the position of such a line, (2) measure the amount of conformity thereto, and (3) translate this result by a prescribed method into an index of correlation. Our first concern, then, would be with the location of the line which best fits the observed data.

As we have previously learned, a relation may be rectilinear or curvilinear; hence, the best-fitting line may be straight or curved. In this section, we restrict ourselves to those relations which are rectilinear, or approximately so, and which can therefore be legitimately represented by a straight line. In other words, our concern will be with those relations in which a constant change in one variable produces, on the average, a constant change in the other paired variable. Such a relation has already been illustrated in Figure 10.2.2, which plots the size of family and the size of farm in China: family size increases by one person on the average as farm size increases by ten acres.

Here, for purposes of exposition, let us use some very simple data which exemplify the relation between income and social status score

(Table 10.3.1a). It is reasonable to designate social status as the dependent Y-variable, since status appears to adjust itself to income in the long run, rather than the reverse. Once cases have been plotted to form a scatter diagram (Figure 10.3.1), we wish to draw the most representative straight line.

Table 10.3.1a

Income and Social Status Score, Selected Families

Income in Thousands of Dollars (X)	Social Status Score (Y)
3	3
7	5
11	7
14	6
15	9
$\bar{X} = 10$	$\bar{Y} = 6$

Source: Suggested by W. Lloyd Warner's studies in social class.

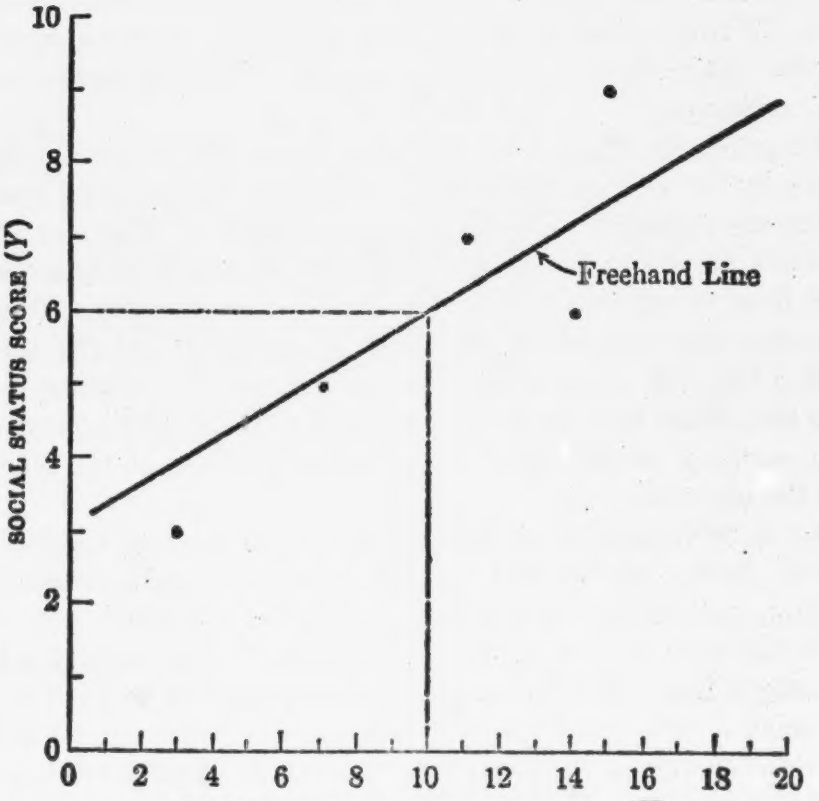


FIGURE 10.3.1 *Freehand Trend Line on Scatter Diagram, Income and Social Status*

Freehand Regression Line. It would of course be possible to draw freehand what appears by inspection to be the best-fitting trend line. If we had drawn such a line and were then asked to defend its location, we would spontaneously reply that we had drawn it "right through the middle" — that is, as close to all the points on the average as possible, cutting the swarm lengthwise into two equal bands. It was drawn down the middle because we intuitively felt that this line was the best average for *all* of the observed points. It thus serves as a running estimate of the *Y*-values, which takes into account every observed value of *X*. If we had been asked to guess a person's social status score from his income, say $10,000, we would not guess a *Y*-value outside, or even close to the edge, of the region of scatter, but rather a central value within the swarm above the designated *X*-value. This would be a status score of 6. Such *estimated values of Y* correspond to the observed *X*'s and constitute the aforementioned trend line. Naturally, the more closely the observed points hug the line of estimation — the less they digress from it — the smaller the error in prediction of *Y* from *X*. If income were the sole determinant of social status, the observed points would necessarily all be on the line, without error; if social status alone determined income, the points would again fall on the line without residue.

Now, what is the significance for correlation of these departures from the hypothetical regression line? To the statistical analyst it indicates that there are *other* factors besides income which determine social status. The operation of these other, unknown factors disturbs our prediction. Hence, they reduce the correlation between the two given variables.

Values Expressed as Deviations from Mean. In Table 10.3.1a and the scatter diagram derived from it, the variates are given in their original form. But for several reasons, not fully amplified here, it is more meaningful to compare the variables after they have been transformed into deviations from their respective means, $\bar{X}$ and $\bar{Y}$.

When comparing values, as we do in correlation, the student must accustom himself to thinking in terms of deviations from the mean instead of the raw data. Like so many other statistical procedures which may at first seem needlessly indirect and complex, this too conforms to common sense. We do not ordinarily compare raw test scores, salaries, birth rates; but rather consider them as "high" or "low" — that is, above or below the average or norm to which we are accustomed. Still less can we compare units of different categories, such as incomes and birth rates in their raw form; we associate income below average with family size that exceeds the average. Thus, we see that the mean is the appropriate point of origin — the natural standard — whenever we compare two sets of data in respect to their affinity with each other.

The items in Table 10.3.1b are therefore expressed as deviations from their respective means, and then plotted in that form (Figure 10.3.2a). In graphing deviations, we must of course be able to accommodate negative values, and consequently we use four quadrants, corresponding to the four combinations of signs.

Table 10.3.1b

Incomes and Social Status Scores as Deviations from $\bar{X}$ and $\bar{Y}$

INCOME IN THOUSANDS OF DOLLARS (x)	SOCIAL STATUS SCORE (y)
−7	−3
−3	−1
1	1
4	0
5	3
0	0

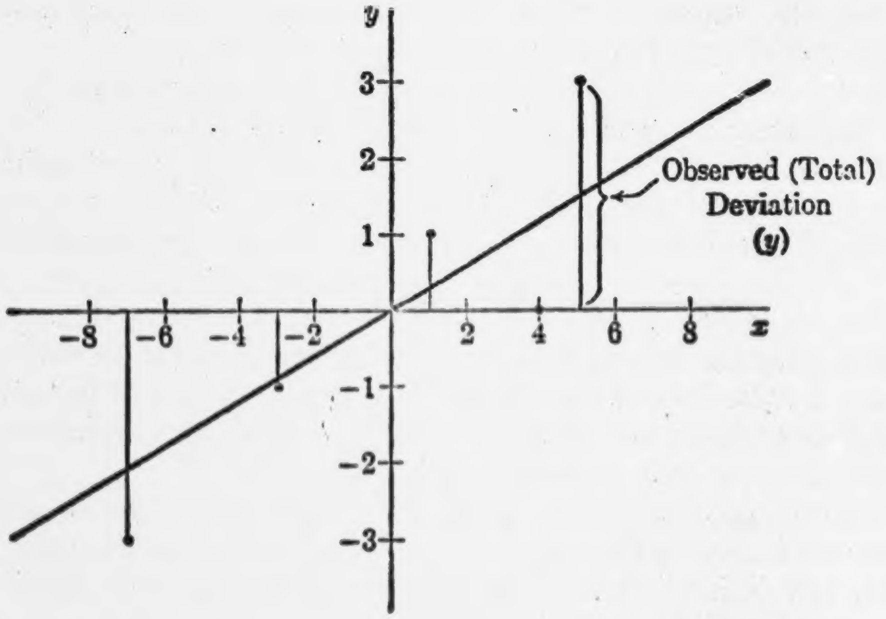


FIGURE 10.3.2a *Scatter Diagram, Observed Deviations* (y)

It should be observed that the configuration of plotted deviations is identical with the scatter of original values; only the scale·markers have been changed. All we did was to shift the zero origin to the intersection of the means.

Explained Deviations. The freehand trend line, drawn straight through the middle, is obviously an average for all of the observed points, a continuous succession of norms, as it were, adjusted to the entire series of cases. Since it consists of the deviations which one would expect on the average, the values on the sloping trend line are variously called the "expected," "predicted," or "estimated" values. The expected deviations are graphically plotted (Figure 10.3.2b) by vertical lines extending

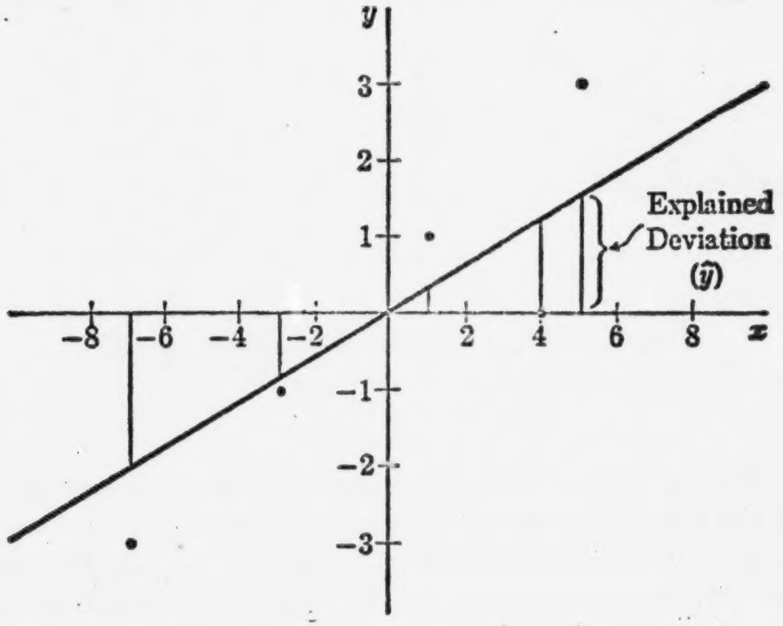


Figure 10.3.2b *Explained Deviations* ($\hat{y}$)

from the x-axis (which coincides with the mean of the Y-values) to the points on the trend line. These are the deviations in social status scores (Y) which one would expect, or predict, *if income* (X) *were the only variable determining them.* We therefore state, in statistical language, that these deviations, which may be either larger or smaller than those observed, are *explained* by income, or attributable to the income factor. Consequently, these deviations have come to be known as ***explained deviations.***

Unexplained Residuals. Since there are other factors that also influence social status (such as occupation, nativity, etc.), the actual, or observed, values do not coincide with the expected values. The force of these unidentified factors is measured by the discrepancies between the observed (O) and the expected (E) deviations. The smaller these discrepancies, the weaker must be the force of the extraneous factors; the larger these discrepancies, the stronger these forces must be. The discrepancies are sometimes called *residuals*, and in terms of the given data remain *un-*

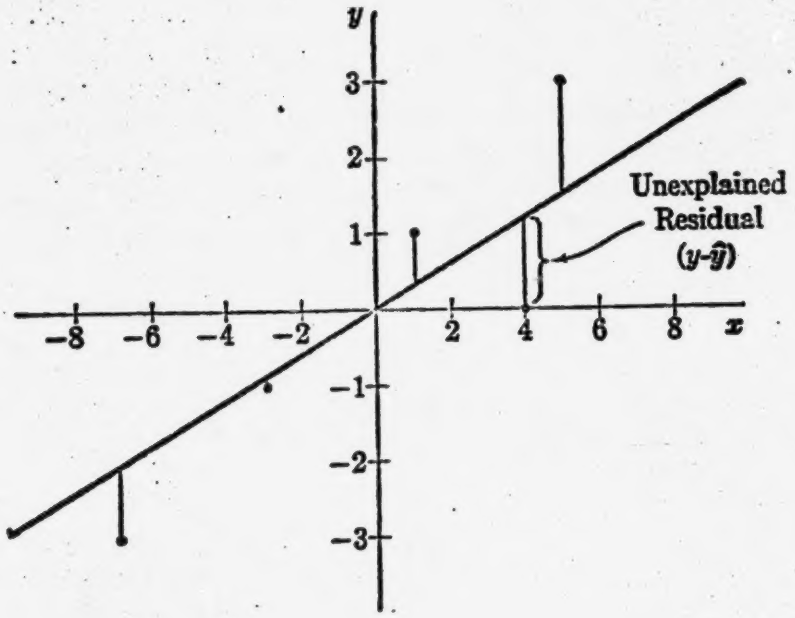


Figure 10.3.2c *Unexplained Residuals* $(y - \hat{y})$

explained; that is, they must be attributed not to income, but to unknown factors on the identity of which we may only speculate, since additional data are not available in the problem as set up. Hence, they are called *unexplained residuals,* measured by the respective distances from the trend line to the individual observed points in the swarm.

From inspection of the foregoing charts, we must conclude that the closer the correspondence between the explained and observed deviations (i.e., the smaller the residuals) the higher the degree of correlation. It would therefore be logical to measure the degree of correlation according to the degree of correspondence between the estimated and observed deviations; or by the proportion of the aggregate observed deviation that is explained. This is the fundamental principle of the measurement of correlation and is always implicit in its calculation. To be sure, the measurement is not carried out informally by freehand regression lines or by graphic readings obtained with a ruler or a pair of dividers; still, it remains essentially a comparison, or ratio, between the explained and total deviations.

Since these concepts play such an important part in the subsequent presentation, it will be well to summarize their meaning precisely:

(1) *Total (Observed) Deviation:* the deviation of an observed value from the mean of the series.

(2) *Explained Deviation:* the deviation of an expected (regression) value from the mean of the series.

(3) *Unexplained Residual:* the discrepancy between the total and the explained deviation. This is, of course, the difference between (1) and (2) above.

The Measurement of Linear Correlation

A serious disadvantage of freehand graphs is that they rest on personal judgment. Without a standard operating procedure, it is unlikely that two workers would ever locate the trend line in exactly the same position. Obviously, a line that is used to measure correlation in the manner suggested above must always meet the same specifications; otherwise the results will be lacking in the reliability which is essential to scientific procedure.

Such standard specifications have been formulated as follows: the line is so located as to make the sum of the vertical residuals around that line equal to zero — that is, to make the sum of the positive and negative discrepancies balance one another. Thus, such a line represents the scatter around it, as the mean represents its array. And like the mean, it makes the sum of the squared deviations a minimum — it conforms to the *principle of least squares.** Because of these properties it is labeled

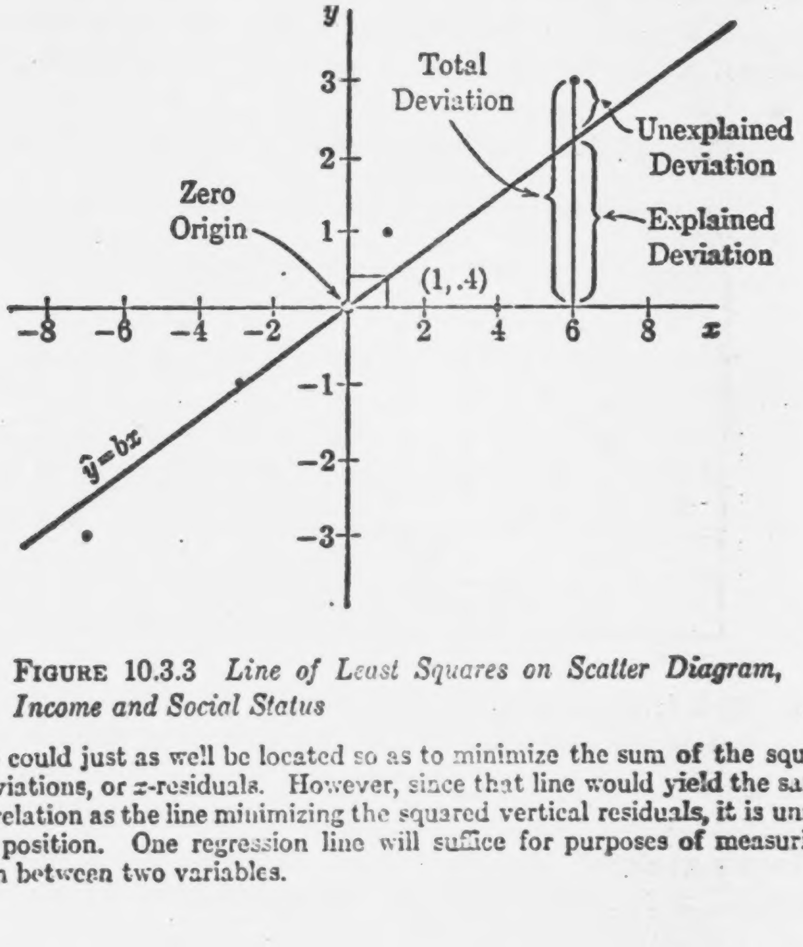


FIGURE 10.3.3 *Line of Least Squares on Scatter Diagram, Income and Social Status*

the line of *best fit;* by the criterion of least squares, it fits the scatter better than does any other straight line.

---

* The line could just as well be located so as to minimize the sum of the squared *horizontal* deviations, or x-residuals. However, since that line would yield the same measure of correlation as the line minimizing the squared vertical residuals, it is unnecessary to fix its position. One regression line will suffice for purposes of measuring linear correlation between two variables.

The tracing of this mathematical line of best fit is simplified by again plotting cases as deviations from means. For, under these conditions, the line of least squares will always conveniently pass through the zero origin, which lies at the intersection of the means. Consequently, it may be readily plotted once its *slope* on the x-axis has been determined.

Computation of Slope. The calculation of this slope is according to the following equation, presented here without explanation:

$$b_{yx} = \frac{\Sigma xy}{\Sigma x^2} \qquad (10.3.1)$$

where b_{yx} = slope of line of expected y's on observed x's
(read: "b, y on x")
Σxy = sum of products of paired x- and y-deviations
(read: "sum of cross-products")

This ratio is the slope of the sought-for best-fitting line; it is the average change in Y per unit change in X. Accordingly, to find this average change in social status per unit change in income, we compute the cross-products and squared x-deviations, and form the ratio between their respective sums. Performing these operations (Table 10.3.2), we obtain $b_{yx} = .4$. This is the average rate of increase in social status per unit of income; that is, for every $1,000 increase in income, there is an average increase of .4 status points. Because this average fixes the inclination of the regression line to the axis of the independent variable, it has come to be known as *slope*. Once determined, it enables us to plot the line of least squares and proceed to the precise measurement of correlation.

Table 10.3.2 *Calculation of* b_{yx}

X	Y	x	y	x^2	xy
3	3	−7	−3	49	21
7	5	−3	−1	9	3
11	7	1	1	1	1
14	6	4	0	16	0
15	9	5	3	25	15
$\Sigma = 50$	30	0	0	100	40

$$b_{yx} = \frac{\Sigma xy}{\Sigma x^2} = \frac{40}{100} = .4$$

Plotting the Regression Line. Since a straight line is determined by any two points, we plot two points known to lie on the least-squares line and run a straight line through them. For one of these, we pick up the point of origin (intersection of axes); the other point is most conveniently obtained by calculating the expected value of y when $x = 1.00$. This will of course be equal to b_{yx}. Accordingly, we draw a line through the zero origin and a point plotted at the intersection of 1.00 and .4. This line is plotted in Figure 10.3.3.

It is this line that supplies the explained deviations ($\hat{y}$'s) corresponding to the observed x's, since the unbroken line of best fit necessarily consists of all possible explained deviations. From it, we may read the approximate magnitudes of the explained deviations; but they may be more precisely determined by application of the slope formula:

$$\hat{y} = b_{yx}(x) \tag{10.3.2}$$

where $\hat{y}$ = expected deviation in Y for any given x (read "y-circumflex"). Substituting the observed x-deviations in this formula, and solving for the $\hat{y}$'s, we obtain the results shown in Table 10.3.3. These $\hat{y}$'s are the y's that would have been observed if X and Y were perfectly correlated. But these $\hat{y}$'s were not observed; hence, we must determine how close they come to those observed, in order to fix the degree of correlation.

Table 10.3.3

Complete Set of $\hat{y}$-Values for Observed x's

b_{yx}	x	$\hat{y}$
.4	−7	−2.8
.4	−3	−1.2
.4	1	.4
.4	4	1.6
.4	5	2.0
	0	0.0

Coefficient of Determination. We have thus come to the final stage in the measurement of correlation. This measurement proceeds on the principle that the more nearly the explained deviations approach the total deviations, obviously the greater is the proportion of variation explained and the higher the degree of correlation. As a statistical operation, the conversion to a single index consists of summing the squared explained deviations, and expressing this *explained variation* as a proportion of the *total variation* to be explained. Symbolically,

$$r^2 = \frac{\text{Explained Variation}}{\text{Total Variation}} = \frac{\Sigma \hat{y}^2}{\Sigma y^2} \tag{10.3.3}$$

This ratio, be it noted, is symbolized by r^2. It is the fraction of the total variation in the dependent Y-variable which is determined by the independent X-variable; hence, it is called the *Coefficient of Determination.* Since it is impossible to explain more than the total variation, r^2 can never be greater than unity, and will practically always be less than one.

Had the simple algebraic deviations around the mean been employed as measures of variation, they would have summed to zero — a very impractical situation which would not have permitted any ratios at all. Hence we turn to squared deviations, in compliance with the principle of least squares.

But in order to circumvent this more complicated computation, would it not have been possible to take the simple arithmetic deviations from the median? This was, in principle, the solution of Francis Galton who originally propagated the concepts of regression and correlation in his famous work *Natural Inheritance* (1889). However, the mean has properties not possessed by the median which make it more useful in correlation computations. And once the mean has been chosen, there is no alternative but to adopt the least-squares principle for its implementation.

The complete computation of r^2 is illustrated in Table 10.3.4, where

Table 10.3.4 *Computation of* r^2

x	y	b	x	$\hat{y}$	$\hat{y}^2$	y^2
−7	−3	(.4)	(−7)	−2.8	7.84	9
−3	−1	(.4)	(−3)	−1.2	1.44	1
1	1	(.4)	(1)	.4	.16	1
4	0	(.4)	(4)	1.6	2.56	0
5	3	(.4)	(5)	2.0	4.00	9
				0.0	16.00	20

$$r^2 = \frac{16}{20} = .80$$

the explained and observed deviations are squared and summed to give the explained and total variation, respectively. Upon dividing the total variation, 20, into the explained variation, 16, we obtain $r^2 = .80$. This is the final measure of the degree of association between the two variables: it reveals that 80 per cent of the variation in Y is accounted for by its linear dependence on X.

Reversibility of r^2. We could have computed the regression of X on Y as easily as the regression of Y on X, and thereby determined the proportion of X-variation explained by Y. In fact, such a result would appear essential to a complete statement of correlation. Why compute r^2_{yx} and ignore r^2_{xy}? The answer is that it has not been ignored, although we did not compute it independently. It is simply unnecessary to fit both regression lines for the reason that $r^2_{yx} = r^2_{xy}$. In short, r^2 is reversible. For example, knowing that $r^2_{yx} = .80$, we may state not only that income explains 80 per cent of the variation in social status but the reverse as well, namely, that social status accounts for 80 per cent of the variation in income. Since, from a purely statistical standpoint, each explains the other to the same degree, the subscript is usually not attached to r^2.

Coefficient of Non-determination. Since the difference between total variation and explained variation is necessarily equal to the unexplained variation, it follows that the unexplained proportion is simply the difference between unity and r^2, a quantity appropriately termed the *Coefficient of Non-determination.* We may write it as follows:

$$1 - r^2 = 1 - \frac{\text{Explained Variation}}{\text{Total Variation}}$$

We could have obtained this quantity directly by measuring the residuals around the regression line, squaring and summing them, and expressing this sum as a proportion of total variation. In fact, this is the operational meaning of $1 - r^2$. But whether calculated directly from the residuals, or indirectly via r^2, this coefficient may always be construed as the proportion of variation in the dependent variable left linearly unexplained by the independent variable. It thereby gauges the strength of the unidentified factors.

Pearsonian Product-Moment r. Conventionally, it is r rather than r^2 that is selected and quoted to indicate the degree of correlation between the two variables. Why is this so, particularly since r^2 seems to have all of the prerequisites for a satisfactory index of correlation: it ranges between zero and unity; it has a simple, comprehensible meaning? Nevertheless, unless another index is specifically named, it is assumed that *the* correlation coefficient refers to *Pearsonian r.* How is this convention explained?

Before formulating an answer to that question, we must first acquaint ourselves with the distinctive properties of r. Accordingly, we return once again to the scatter diagram and the line of best fit. Our procedure will be exactly the same as that followed previously, excepting that the data will now be plotted as standard deviates (σ-values), instead of the observed deviations (Figure 10.3.4). Again the configuration of points is left unaffected by the transformation: the sole alteration is that the scale unit is now the standard deviation rather than the original raw unit of measure.

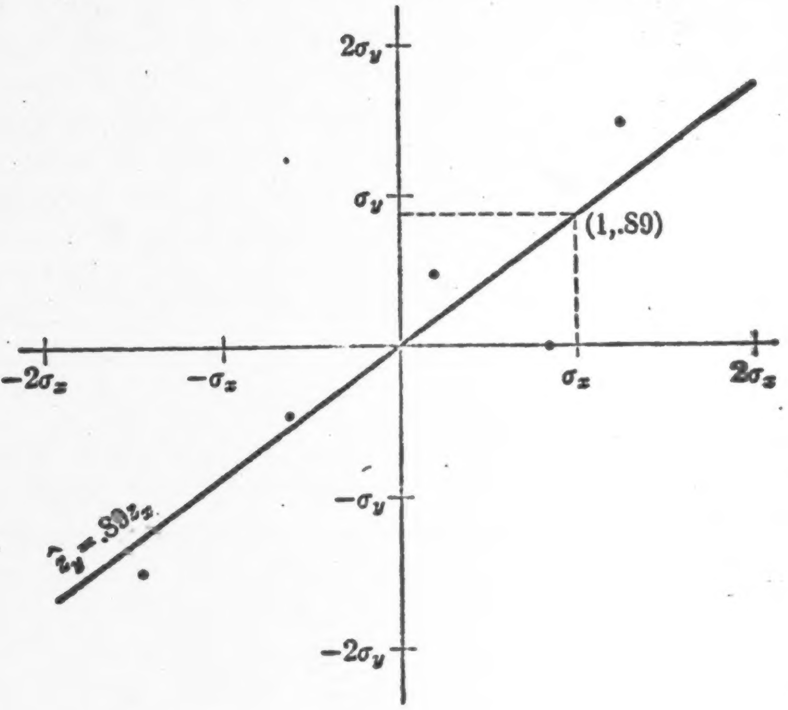


Figure 10.3.4 *Scatter Diagram of Standard Deviates*

Table 10.3.5

Deviations as Standard Deviates

Deviations		Standard Deviates	
x	y	$\frac{x}{\sigma}$	$\frac{y}{\sigma}$
−7	−3	−1.56	−1.50
−3	−1	− .67	− .50
1	1	.22	.50
4	0	.89	.00
5	3	1.12	1.50
0	0	0.00	0.00

$\sigma_x = 4.47$
$\sigma_y = 2.00$

To establish the line of best fit through this scatter of σ-points, we first calculate its slope as before, except that now we operate on the standard deviates instead of the raw deviations. Consequently, the slope would now be written in standard form:

$$\frac{\sum\left(\frac{x}{\sigma}\right)\left(\frac{y}{\sigma}\right)}{\sum\left(\frac{x}{\sigma}\right)^2}$$

instead of:

$$\frac{\Sigma xy}{\Sigma x^2}$$

Since $\frac{x}{\sigma}$ is usually symbolized z, the slope formula may be simply written as:

$$r_{yx} = \frac{\Sigma z_x z_y}{N} \qquad (10.3.4)$$

and symbolized r_{yx} instead of b_{yx}. This is the culmination of our development. We are now in a position to see that the slope of the regression line through the scatter of σ-points is identical with the square root of explained variation: that is, $r = \sqrt{r^2}$. For the example above (Table 10.3.4), $\sqrt{.80} = \pm.89$.

To compute r directly, we would proceed as in Table 10.3.6, where paired standard deviates are multiplied together and those cross-products are summed and averaged. The mean of the cross-products is r.

Table 10.3.6

Pearsonian r as Mean of Cross-Products of Standard Deviates

z_x	z_y	$z_x z_y$
−1.56	−1.50	2.340
− .67	− .50	.335
.22	.50	.110
.89	.00	.000
1.12	1.50	1.680
		4.465

$$r = \frac{\Sigma z_x z_y}{N} = \frac{4.46}{5} = .89$$

We can now understand why r was designated by Karl Pearson "product-moment correlation coefficient." From the composition of Formula 10.3.4 above, it is clear that r is strictly an arithmetic mean; it is a sum divided by the number of items in that sum. Now, the terms "mean" and "moment" are used interchangeably in mathematical statistics; hence, Pearson labeled the mean of the standard cross-products the product-moment correlation coefficient.

Viewed in this light, r may be construed as the mean change in Y for every unit change in X, or the reverse, always assuming measures in standard form, i.e., expressed in terms of sigma units. Thus, if a given income deviates by 1σ from its mean, we expect the associated status scores to deviate on the average from their mean by $.89\sigma$. By virtue of this principle, we may estimate Y (in standard form) for any given value of X, exactly as we applied b_{yx} to obtain the expected y-deviations. We have only to apply the formula:

$$\hat{z}_y = r(z_x) \qquad (10.3.5)$$

where $\hat{z}_y$ = estimated σ-value in Y
z_x = observed σ-value in X

Substituting our data in this formula, we get the estimated values shown in Table 10.3.7. Thus, when income is 1.12 in standard form, social status is estimated to be 1.00; when income is -1.56σ, social status would be -1.39σ. Perfect correlation between two series would prevail whenever the paired values are identical sigma distances from their respective means — for example, if the person with a sociology grade of, say, 1.8σ above the mean were located 1.8σ above the mean of the history grades, and all other observations in the two series were also perfectly paired. Because r is reversible, we could have applied r to the observed z_y's and thereby estimated the corresponding z_x's.

Table 10.3.7 *Estimated z_y from Observed z_x*

z_x	$(r)z_x$	=	$\hat{z}_y$
-1.56	$(.89)(-1.56)$	=	-1.39
$-\ .67$	$(.89)(-\ .67)$	=	$-\ .60$
.22	$(.89)(\ .22)$	=	.20
.89	$(.89)(\ .89)$	=	.79
1.12	$(.89)(\ 1.12)$	=	1.00

Comparison of r and r^2. On the surface, the difference between r and r^2 appears very trivial: a simple detail of exponent, with both values easily convertible into the other, after one has been computed. Nevertheless, this seemingly innocent difference cannot be so lightly dismissed, for the two r's respectively focus on two distinct but interrelated aspects of covariation — a distinction which the design of the present chapter has deliberately sought to portray.

Either one could have been mechanically derived from the other by way of only one of the two procedures. However, an r^2 derived from r would never convey to the student the intent and meaning of explained variation which r^2 measures. The "square of the slope" would mean nothing to him. Nor would an r, converted arithmetically from r^2, convey the intent and meaning of the slope. The "square root of explained variation" could not be visualized as slope. Although r and r^2 are inter-dependent, r^2 measures the over-all proportion of the total variation of one variable that is associated with, or explained by, the other.

On the other hand, r measures the dynamic aspect of this relation, measuring the rate of change in one variable relative to the other, as has been demonstrated in the paradigms above. Because of this conceptual distinction, we may say that r is primarily a predictive device to forecast, for example, the expected level of performance on one variable from observed performance on another. As such, it is probably more likely to be used by educators, psychologists, and others interested in personal prediction, rather than by sociologists. On the other hand, r^2 is a summarizing measure weighing the influence, or force, exerted by one variable on the other.

Since r is slope, it manifestly must have a direction: predominantly up or down, according to whether the variables are positively or negatively related. It follows that the direction of the slope reflects the type of relation, which is then symbolized by a plus or minus sign. The synoptic measure r^2, however, carries no sign since it expresses a proportion of a total variation.

It is now clear that r and r^2 are not interchangeable; nor should they be cheaply derived from each other until their structural meanings are thoroughly understood. Since r is always larger than r^2, it could be deliberately or unwittingly used to exaggerate the strength of association and thereby mislead the reader: for example, $r = .5$ may seem to signify a reasonably strong association, but $r^2 = .25$ indicates that only 25 per cent of the variation in either variable is accounted for by the other. When the emphasis is on the strength of the over-all relationship between two variables, as is frequently the case in sociological studies, r^2 is the pertinent statistic. For example, the finding that 50 per cent of the variation in area delinquency may be attributed to the economic factor still leaves something unexplained; yet it represents a tangible gain in understanding a phenomenon that until a century ago was laid at the door of demons, heredity, or free will.

In sum, each measure has its very distinctive connotations and appropriate uses. Nevertheless, r enjoys a near monopoly over r^2. This may be due partly to the inertia of tradition, which will perhaps be dissipated when workers become more sensitive to the nuances of quantitative reasoning.

Computing Formulas: Ungrouped Data. The extended analytic operations previously given were designed to expose the logic and intent of the concepts of r and r^2, so that they could be discriminatingly applied in social research. But such elaborate calculations need not be carried out if the sole object is to determine the mere numerical value of r or r^2 in a given practical problem. There are handbook working procedures that obviate the need to compute slope, the explained deviations, the unexplained residuals, and other quantities involved in the formulas given above. Of course, such mechanical computing methods are not interpretive tools; they are merely recipes that can be followed by any intelligent clerk or machine.

The computing programs here given begin with the product-moment formula, which may be rewritten for somewhat simpler calculation:

$$r = \frac{\Sigma xy}{\sqrt{\Sigma x^2 \Sigma y^2}} \qquad (10.3.6)$$

And it may be still more conveniently rewritten, particularly when a desk calculator is available, in terms of the original, untransformed values:

$$r = \frac{N\Sigma XY - \Sigma X \Sigma Y}{\sqrt{[N\Sigma X^2 - (\Sigma X)^2][N\Sigma Y^2 - (\Sigma Y)^2]}} \qquad (10.3.7)$$

This is the computing formula for r, ungrouped data. It intends to eliminate many of the vexatious arithmetic details that are sure to arise when the operation is carried out on the raw deviations around means.

This formula, be it noted, requires only five sums and correspondingly five work columns. Two of these columns comprise the original tabulation of X and Y, so that only three more are needed in order to proceed with the calculation of r. Substituting sums in the formula, we again obtain $r = .89$.

Table 10.3.8 *Worktable for Calculation of r from Ungrouped Values*

X	Y	X^2	Y^2	XY
3	3	9	9	9
7	5	49	25	35
11	7	121	49	77
14	6	196	36	84
15	9	225	81	135
$\Sigma = 50$	30	600	200	340

$$\begin{aligned} r &= \frac{5(340) - (50)(30)}{\sqrt{[5(600) - (50)^2][5(200) - (30)^2]}} \\ &= \frac{1700 - 1500}{\sqrt{(500)(100)}} \\ &= \frac{200}{224} \\ &= .89 \end{aligned}$$

CAUSATION AND THE LAW

One of the most important elements in the attribution of legal responsibility for negligence in cases involving torts is the causal relation between defendant's fault and plaintiff's injury. In the law, however, as Harper and James point out in the first of the following two readings, there has developed the concept of "proximate" or "legal" cause, which differs from the scientific meaning of cause in that it limits and restricts liability. The result, they note, has been a widely recognized confusion.

It would be insufficient, however, merely to state that proximate cause is a legal device to restrict liability; the student should also note the fundamental difference between the ways in which the lawyer and scientist approach causation. The selection from Hart and Honore discusses this difference. According to the authors, the kinds of inquiries made by scientists are explanatory while those of the lawyer are attributive. The scientist's concern is to isolate, as far as possible, all antecedents of a given consequence, for example, in the construction of experiments. The lawyer, on the other hand, pursues a much less ambitious goal, focusing on whether a particular defendant's conduct was culpable. In other words, there may well be multiple causes of an event, and it may be the task of science to disentangle this complex network; but for the purposes of the law, it is sufficient to establish the nature of the defendant's contribution.

From H. L. A. Hart and A. M. Honore, Causation in the Law (Oxford: The Clarendon Press, 1959) pp. 26-30; 64-78. Reprinted by permission.

From Fowler V. Harper and Fleming James, Jr., The Law of Torts (Boston: Little, Brown, & Co., Inc., 1956) pp. 1108-1121. Reprinted by permission.

H.L.A. Hart and A.M. Honore, Causation in the Law
(Oxford: Oxford University Press, 1959), pp. 64-78.

'To consequences no limit can be set': 'Every event which would not have happened if an earlier event had not happened is the consequence of that earlier event.' These two propositions are not equivalent in meaning and are not equally or in the same way at variance with ordinary thought. They have, however, both been urged sometimes in the same breath by the legal theorist[1] and the philosopher: they are indeed sometimes said by lawyers to be 'the philosophical doctrine' of causation. It is perhaps not difficult even for the layman to accept the first proposition as a truth about certain physical events; an explosion may cause a flash of light which will be propagated as far as the outer nebulae; its effects or consequences continue indefinitely. It is, however, a different matter to accept the view that whenever a man is murdered with a gun his death was the consequence of (still less an 'effect' of or 'caused by') the manufacture of the bullet. The first tells a perhaps unfamiliar tale about unfamiliar events; the second introduces an unfamiliar, though, of course, a possible way of speaking about familiar events. It is not that this unrestricted use of 'consequence' is unintelligible or never found; it is indeed used to refer to bizarre or fortuitous connexions or coincidences: but the point is that the various causal notions employed for the purposes of explanation, attribution of responsibility or the assessment of contributions to the course of history carry with them implicit limits which are similar in these different employments.

It is, then, the second proposition, defining consequence in terms of 'necessary condition', with which theorists are really concerned. This proposition is the corollary of the view that, if we look into the past of any given event, there is an infinite number of events, each of which is a necessary condition of the given event and so, as much as any other, is its cause. This is the 'cone'[2] of causation, so-called because, since any event has a number of simultaneous conditions, the series fans out as we go back in time. The justification, indeed only partial, for calling this 'the philosophical doctrine' of causation is that it resembles Mill's doctrine that 'we have no right to give the name of cause to one of the conditions exclusive of the others of them'. It differs from

---

[1] Lawson, *Negligence in the Civil Law*, p. 53.
[2] Glanville Williams, *Joint Torts and Contributory Negligence*, p. 239.

Mill's view in taking the essence of causation to be 'necessary condition' and not 'the sum total'[1] of the sufficient conditions of an event.

Legal theorists have developed this account of cause and consequence to show what is 'factual', 'objective', or 'scientific' in these notions: this they call 'cause in fact' and it is usually stressed as a preliminary to the doctrine that any more restricted application of these terms in the law represents nothing in the facts or in the meaning of causation, but expresses fluctuating legal policy or sentiments of what is just or convenient. Moral philosophers have insisted in somewhat similar terms that the consequences of human action are 'infinite': this they have urged as an objection against the Utilitarian doctrine that the rightness of a morally right action depends on whether its consequences are better than those of any alternative action in the circumstances. 'We should have to trace as far as possible the consequences not only for the persons affected directly but also for those indirectly affected and to these no limit can be set.'[2] Hence, so the argument runs, we cannot either inductively establish the Utilitarian doctrine that right acts are 'optimific' or use it in particular cases to discover what is right. Yet, however vulnerable at other points Utilitarianism may be as an account of moral judgment, this objection seems to rest on a mistake as to the sense of 'consequence'. The Utilitarian assertion that the rightness of an action depends on its consequences is not the same as the assertion that it depends on all those later occurrences which would not have happened had the action not been done, to which indeed 'no limit can be set'. It is important to see that the issue here is not the linguistic one whether the word 'consequence' would be understood if used in this way. The point is that, though we could, we do not think in this way in tracing connexions between human actions and events. Instead, whenever we are concerned with such connexions, whether for the purpose of explaining a puzzling occurrence, assessing responsibility, or giving an intelligible historical narrative, we employ a set of concepts restricting in various ways what counts as a consequence. These restrictions colour *all* our thinking in causal terms; when we find them in the law we are not finding something invented by or peculiar to the law, though of course it is for the law to say when and how far it will use them and, where they are vague, to supplement them.

No short account can be given of the limits thus placed on 'consequences' because these limits vary, intelligibly, with the variety of causal connexion asserted. Thus we may be tempted by the generalization that consequences must always be something intended or foreseen or at least foreseeable with ordinary care: but counter-examples spring up from many types of context where causal statements are made. If smoking is shown to cause lung cancer this discovery will permit us to describe past as well as future cases of cancer as the effect or consequence of smoking even though no one foresaw or had reasonable grounds to suspect this in the past. What is common and commonly appreciated and hence foreseeable certainly controls the scope of consequences in certain varieties of causal statement but not in all. Again the voluntary intervention of a second person very often constitutes the limit. If a guest sits down at a table laid with knife and fork and plunges the knife into his hostess's breast, her death is not in any context thought of as caused by, or the effect or result of the waiter's action in laying the table; nor would it be linked with this action as its consequence for any of the purposes, explanatory or attributive, for which we employ causal notions. Yet as we have seen there are many other types of case where a voluntary action or the harm it does are naturally attributed to some prior neglect of precaution as its consequence. Finally, we may think that a simple answer is already supplied by Hume and Mill's doctrine that causal connexion rests on general laws asserting regular connexion; yet, even in the type of case to which this important doctrine applies, reference to it alone will not solve our problem. For we often trace a causal connexion between an antecedent and a consequent which themselves very rarely go together: we do this when the case can be broken down into intermediate stages, which themselves exemplify different generalizations, as when we find that the fall of a tile was the cause of someone's death, rare though this be. Here our problem reappears in the form of the question: When can generalizations be combined in this way?

---

[1] Mill, Book III, chap. v, s. 2.

[2] Ross, *The Right and the Good*, p. 36.

We shall examine first the central type of case where the problem is of this last-mentioned form. Here the gist of the causal connexion lies in the general connexion with each other of the successive stages; and is not dependent on the special notions of one person providing another with reasons or exceptional opportunities for actions. This form of causal connexion may exist between actions and events, and between purely physical events, and it is in such cases that the words 'cause' and 'causing' used of the antecedent action or event have their most obvious application. It is convenient to refer to cases of the first type where the consequence is harm as cases of 'causing harm', and to refer to cases where harm is the consequence of one person providing another with reasons or opportunities for doing harm as cases of 'inducing', 'advising', or 'occasioning' harmful acts.[1] In cases of the first type a voluntary act, or a conjunction of events amounting to a coincidence, operates as a limit in the sense that

[1] Footnote omitted.

events subsequent to these are not attributed to the antecedent action or event as its consequence even though they would not have happened without it. Often such a limiting action or coincidence is thought of and described as 'intervening': and lawyers speak of them as 'superseding' or 'extraneous' causes 'breaking the chain of causation'. To see what these metaphors rest on (and in part obscure) and how such factors operate as a limit we shall consider the detail of three simple cases.

(i) A forest fire breaks out, and later investigation shows that shortly before the outbreak A had flung away a lighted cigarette into the bracken at the edge of the forest, the bracken caught fire, a light breeze got up, and fanned the flames in the direction of the forest. If, on discovering these facts, we hesitate before saying that A's action caused the forest fire this would be to consider the alternative hypothesis that in spite of appearances the fire only succeeded A's action in point of time, that the bracken flickered out harmlessly and the forest fire was caused by something else. To dispose of this it may be necessary to examine in further detail the process of events between the ignition of the bracken and the outbreak of fire in the forest and to show that these exemplified certain types of continuous change. If this is shown, there is no longer any room for doubt: A's action *was* the cause of the fire, whether he intended it or not. This seems and is the simplest of cases. Yet it is important to notice that even in applying our general knowledge to a case as simple as this, indeed in regarding it as simple, we make an implicit use of a distinction between types of factor which constitute a limit in tracing consequences and those which we regard as mere circumstances 'through' which we trace them. For the breeze which sprang up after A dropped the cigarette, and without which the fire would not have spread to the forest, was not only subsequent to his action but entirely independent of it: it was, however, a common recurrent feature of the environment, and, as such, it is thought of not as an 'intervening' force but as merely part of the circumstances in which the cause 'operates'. The decision so to regard it is implicitly taken when we combine our knowledge of the successive stages of the process and assert the connexion.

It is easy here to be misled by the natural metaphor of a causal 'chain', which may lead us to think that the causal process consists of a series of single events each of which is dependent upon (would not have occurred without) its predecessor in the 'chain' and so is dependent upon the initiating action or event. In truth in any causal process we have at each phase not single events but complex sets of conditions, and among these conditions are some which are not only subsequent to, but independent of the initiating action or event. Some of these independent conditions such as the evening breeze in the example chosen, we classify as mere conditions in or on which the cause operates; others we speak of as 'interventions' or 'causes'. To decide how such independent elements shall be classified is also to decide how we shall combine our knowledge of the different general connexions which the successive stages exemplify, and it is important to see that nothing *in* this knowledge itself can resolve this point. We may have to go to science for the relevant general knowledge before we can assert with proper confidence that A's action did cause the fire, but science, though it tells us that an air current was required, is silent on the difference between a current in the form of an evening breeze and one produced by someone who deliberately fanned the flames as they were flickering out in the bracken. Yet an air current in this form is not a 'condition' or 'mere circumstance' through which we can trace the consequence; its presence would force us to revise the assertion that A caused the fire. Conversely if science helped us to identify as a necessary factor in producing the fire some condition or element of which we had previously been totally ignorant, e.g. the persistence of oxygen, this would leave our original judgment undisturbed if this factor were a common or pervasive feature of the environment or of the thing in question. There is thus indeed an important sense in which it is true that the distinctions between cause and conditions is not a 'scientific' one. It is not determined by laws or generalizations concerning connexions between events.

When we have assembled all our knowledge of the factors involved in the fire, the residual question which we then confront (the attributive question) may be typified as follows: Here is A's action, here is the fire: can the fire be attributed to A's action as its consequence given that there is also this third factor (the breeze or B's intervention) without which the fire would not have happened? It is plain that, both in raising questions of this kind and in answering them, ordinary thought is powerfully influenced by the analogy between the straightforward cases of causal attribution (where the elements required for the production of harm in addition to the initiating action are all 'normal' conditions) and even simpler cases of responsibility which we do not ordinarily describe in causal language at all but by the simple transitive verbs of action. These are the cases of the direct manipulation of objects involving changes in them or their position: cases where we say 'He pushed it', 'He broke it', 'He bent it.' The cases which we do confidently describe in causal language ('The fire was caused by his carelessness', 'He caused a fire') are cases where no other human action or abnormal occurrence is required for the production of the effect, but only normal conditions. Such cases appear as mere long range or less direct versions or extensions of the most obvious and fundamental case of all for the attribution of responsibility: the case where we can simply say 'He did it.' Conversely in attaching importance to thus causing harm as a distinct ground of responsibility and in taking certain kinds of factor (whether human interventions or abnormal occurrences), without which the initiating action would not have led to harm, to preclude the description of the case in simple causal terms, common sense is affected by the fact that here, because of the manner in which the harm eventuates, the outcome cannot be represented as a mere extension of the initiating action; the analogy with the fundamental case for responsibility ('He did it') has broken down.

When we understand the power exerted over our ordinary thought by the conception that causing harm is a mere extension of the primary case of doing harm, the inter-related metaphors which seem natural to lawyers and laymen, in describing various aspects of causal connexion, fall into place and we can discuss their factual basis. The persistent notion that some kinds of event required in addition to the initiating action for the production of harm 'break the chain of causation' is intelligible, if we remember that though such events actually complete the *explanation* of the harm (and so *make* rather than *break* the causal explanation) they do, unlike mere normal conditions, break the *analogy* with cases of simple actions. The same analogy accounts for the description of these factors as 'new actions' (*novus actus*) or 'new causes', 'superseding', 'extraneous', 'intervening forces': and for the description of the initiating action when 'the chain of causation' is broken as 'no longer operative', 'having worn out', *functus officio*.[1] So too when the 'chain' is held not to be 'broken' the initiating action is said to be still 'potent',[2] 'continuing', 'contributing', 'operative', and the mere conditions held insufficient to break the chain are 'part of the background',[3] 'circumstances in which the cause operates',[4] 'the stage set', 'part of the history'.

(ii) A throws a lighted cigarette into the bracken which catches fire. B, just as the flames are about to flicker out, deliberately pours petrol on them. The fire spreads and burns down the forest. A's action, whether or not he intended the forest fire, was not the cause of the fire: B's was.

The voluntary intervention of a second human agent, as in this case, is a paradigm among those factors which preclude the assimilation in causal judgments of the first agent's connexion with the eventual harm to the case of simple direct manipulation. Such an intervention displaces the prior action's title to be called the cause and, in the persistent metaphors found in the law, it 'reduces' the earlier action and its immediate effects to the level of 'mere circumstances' or 'part of the history.' B in this case was not an 'instrument' through which A worked or a victim of the circumstances A has created. He has, on the contrary, freely exploited the circumstances and brought about the fire without the co-operation of any further agent or any chance coincidence. Compared with this the claim of A's action to be ranked the cause of the fire fails. That this and not the moral appraisal of the two actions is the point of comparison seems clear. If A and B both intended to set the forest on fire, and this destruction is accepted as something wrong or wicked, their moral wickedness, judged by the criterion of intention, is the same. Yet the causal judgment differentiates between them. If their moral guilt is judged by the outcome, this judgment though it would differentiate between them cannot be the source of the causal judgment; for it presupposes it. The difference just is that B has caused the harm and A has not. Again, if we appraise these actions as good or bad from different points of view, this leaves the causal judgments unchanged. A may be a soldier of one side anxious to burn down the enemy's hide-out: B may be an enemy soldier who has decided that his side is too iniquitous to defend. Whatever is the moral judgment passed on these actions by different speakers it would remain true that A had not caused the fire and B had.

---

[1] *Davies* v. *Swan Motor Co.*, [1947] 2 K.B. 291, 318.

[2] *Minister of Pensions* v. *Chennell*, [1947] K.B. 250, 256. Lord Wright (1950), 13 *Mod. L.R.* 3.

[3] *Norris* v. *William Moss & Son Ltd.*, [1954] 1 W.L.R. 46, 351.

[4] *Minister of Pensions* v. *Chennell*, [1947] K.B. 250, 256.

There are, as we have said, situations in which a voluntary action would not be thought of as an intervention precluding causal connexion in this way. These are the cases discussed further below where an opportunity commonly exploited for harmful actions is negligently provided, or one person intentionally provides another with a certain type of reason for wrongdoing. Except in such cases a volunary intervention is a limit past which consequences are not traced. By contrast, actions which in any of a variety of different ways are less than fully voluntary are assimilated to the means by which or the circumstances in which the earlier action brings about the consequences. Such actions are not the outcome of an informed choice made without pressure from others, and the different ways in which human action may fall short in this respect range from defective muscular control, through lack of consciousness or knowledge, to the vaguer notions of duress and of predicaments, created by the first agent for the second, in which there is no 'fair' choice.

In considering examples of such actions and their bearing on causal judgments there are three dangers to avoid. It would be folly to think that in tracing connexions through such actions instead of regarding them, like voluntary interventions, as a limit, ordinary thought has clearly separated out their non-voluntary aspect from others by which they are often accompanied. Thus even in the crude case where A lets off a gun (intentionally or not) and startles B, so that he makes an involuntary movement of his arm which breaks a glass, the commonness of such a reaction as much as its compulsive character may influence the judgment that A's action was the cause of the damage.

Secondly we must not impute to ordinary thought all the fine discriminations that could be made and in fact are to be found in a legal system, or an equal willingness to supply answers to complex questions in causal terms. Where there is no precise system of punishment, compensation or reward to administer, ordinary men will not often have faced such questions as whether the injuries suffered by a motorist who collides with another in swerving to avoid a child are consequences attributable to the neglect of the child's parents in allowing it to wander on to the road. Such questions courts have to answer and in such cases common judgments provide only a general, though still an important indication of what are the relevant factors.

Thirdly, though very frequently non-voluntary actions are assimilated to mere conditions or means by which the first agent brings about the consequences, the assimilation is never quite complete. This is manifested by the general avoidance of many causal locutions which are appropriate when the consequences are traced (as in the first case) through purely physical events. Thus even in the case in which the second agent's rôle is hardly an 'action' at all, e.g. where A hits B, who staggers against a glass window and breaks it, we should say that A's blow made B stagger and break the glass, rather than that A's blow caused the glass to break, though in any explanatory or attributive context the case would be *summarized* by saying that A's action was the cause of the *damage* or that A had caused it.

In the last two cases where B's movements are involuntary in the sense that they are not part of any action which he chose or intended to do, their connexion with A's action would be described by saying that A's blow *made* B stagger or *caused* him to stagger or that the noise of A's shot *made* him jump or *caused* him to jump. This would be true, whether A intended or expected B to react in this way or not, and the naturalness of treating A's action as the cause of the ultimate damage is due to the causal character of this part of the process involving B's action. The same is however true where B's actions are not involuntary movements but A is considered to have made or caused B to do them by less crude means. This is the case if, for example, A uses threats or exploits his authority over B to make B do something, e.g. knock down a door. At least where A's threats are of serious harm, or B's act was unquestionably within A's authority to order, he too has made or forced or (in formal quasi-legal parlance) 'caused' B to act.

Outside the area of such cases, where B's will would be either said not to be involved at all, or to be overborne by A, are cases where A's act creates a predicament for B *narrowing* the area of choice so that he has either to inflict some harm on himself or others, or sacrifice some important interest or duty. Such cases resemble coercion in that A narrows the area of B's choice but differ from it in that this predicament need not be intentionally created. A sets a house on fire (intentionally or unintentionally): B to save himself has to jump from a height involving certain injury, or to save a child rushes in and is seriously burned. Here of course B's movements are not involuntary; the 'necessity' of his action is here of a different order. His action is the outcome of a choice between two evils forced on him by A's action. In such cases, when B's injuries are thought of as the consequence of the fire, the implicit judgment is made that his action was the lesser of two evils and in this sense a 'reasonable' one which he was obliged to make to avoid the greater evil. This is often paradoxically, though understandably, described by saying that here the agent 'had no choice' but to do what he did. Such judgments involve a comparison of the importance of the respective interests sacrificed and preserved, and the final assertion that A's action was the cause of the injuries rests on evaluations about which men may differ.

Finally, the ground for treating some harm which would not have occurred without B's action as the consequence of A's action may be that B acted in ignorance of or under a mistake as to some feature of the situation created by A. Poisoning offers perhaps the simplest example of the bearing on causal judgments of actions which are less than voluntary in this Aristotelian sense. If A intending B's death deliberately poisons B's food and B, knowing this, deliberately takes the poison and dies, A has not caused B's death: if however B does not know the food to be poisoned, eats it and dies A has caused his death, even if he put the poison in unwittingly. Of course only the roughest judgments are passed in causal terms in such cases outside law courts, where fine degrees of 'appreciation' or reckless shutting of the eyes, may have to be discriminated from 'full knowledge'. Yet, rough as these are, they indicate clearly enough the controlling principles.

Though in the foregoing cases A's initiating action might often be described as 'the cause' of the ultimate harm, this linguistic fact is of subordinate importance to the fact that, for whatever purpose, explanatory, descriptive or evaluative, consequences of an action are traced, discriminations are made (except in the cases discussed later) between free voluntary interventions and less than voluntary reactions to the first action or the circumstances created by it.

(iii) The analogy with single simple actions which guides the tracing of consequences may be broken by certain kinds of conjunctions of physical events. A hits B who falls to the ground stunned and bruised by the blow; at that moment a tree crashes to the ground and kills B. A has certainly caused B's bruises but not his death: for though the fall of the tree was, like the evening breeze, in our earlier example, independent of and subsequent to the initiating action, it would be differentiated from the breeze in any description in causal terms of the connexion of B's death with A's action. It is to be noticed that this is not a matter which turns on the intention with which A struck B. Even if A hit B inadvertently or accidentally his blow would still be the cause of B's bruises: he would have caused them though unintentionally. Conversely even if A had intended his blow to kill, this would have been an attempt to kill but still not the cause of B's death. On this legal and ordinary judgments would be found to agree; and most legal systems would distinguish for the purposes of punishment[1] an attempt with a fatal upshot, issuing by such chance or anomalous events, from 'causing death'—the terms in which the offences of murder and manslaughter are usually defined.

Similarly the causal description of the case does not turn on the moral appraisal of A's action or the wish to punish it. A may be a robber and a murderer and B a saint guarding the place A hoped to plunder. Or B may be a murderer and A a hero who has forced his way into B's retreat. In both cases the causal judgment is the same. A had caused the minor injuries but not B's death, though he tried to kill him. A may indeed be praised or blamed but not for causing B's death. However intimate the connexion between causation and responsibility, it does not determine causal judgments in this simple way. Nor does the causal judgment turn on a refusal to attribute grave consequences to actions which normally have less serious results. Had A's blow killed B outright and the tree, falling on his body, merely smashed his watch we should still treat the coincidental character of the fall of the tree as determining the form of causal statement. We should then recognize A's blow as the cause of B's death but not of the breaking of the watch.

The connexion between A's action and B's death in the first case would naturally be described in the language of *coincidence*. 'It was a coincidence: it just happened that, at the very moment when A knocked B down, a tree crashed at the very place where he fell and killed him.' The common legal metaphor would describe the fall of the tree as an 'extraneous' cause. This, however, is dangerously misleading, as an analysis of the notion of coincidence will show. It suggests merely an event which is subsequent to and independent of some other contingency, and of course the fall of the tree has both these features in relation to A's blow. Yet in these respects the fall of

[1] Footnote omitted.

the tree does not differ from the evening breeze in the earlier case where we found no difficulty in tracing causal connexion.[1] The full elucidation of the notion of a coincidence is a complex matter for, though it is very important as a limit in tracing consequences, causal questions are not the only ones to which the notion is relevant. The following are its most general characteristics. We speak of a coincidence whenever (1) the conjunction of two or more events in certain spatial or temporal relations is very unlikely by ordinary standards and (2) is for some reason significant or important, provided (3) that they occur without human contrivance and (4) are independent of each other. It is therefore a coincidence if two persons known to each other in London meet without design in Paris on their way to separate independently chosen destinations; or if two persons living in different places independently decide to write a book on the same subject. The first is a coincidence of time and place ('It just happened that we were at the same place at the same time'), and the second a coincidence of time only ('It just happened that they both decided to write on the subject at the same time').

Use of this general notion is made in the special case when the conjunction of two or more events occurs in temporal and/or spatial relationships which are significant, because, as our general knowledge of causal processes shows, this conjunction is required for the production of some given further event. In the language of Mill's idealized model, they form a necessary part of a complex set of jointly sufficient conditions. In the present case the fall of the tree just as B was struck down within its range satisfies the four criteria for a coincidence which we have enumerated. First, though neither event was of a very rare or exceptional kind, their conjunction would be rated very unlikely judged by the standards of ordinary experience. Secondly, this conjunction was causally significant for it was a necessary part of the process terminating in B's death. Thirdly, this conjunction was not consciously designed by A; had he known of the impending fall of the tree and hit B with the intention that he should fall within its range B's death would not have been the result of any coincidence. A would certainly have caused it. The common-sense principle that a contrived conjunction cannot be a coincidence is the element of truth in the legal maxim (too broadly stated even for legal purposes) that an intended consequence cannot be too 'remote'. Fourthly, each member of the conjunction in this case was independent of the other; whereas if B had fallen against the tree with an impact sufficient to bring it down on him, this sequence of physical events, though freakish in its way, would not be a coincidence and in most contexts of ordinary life, as in the law, the course of events would be summarized by saying that in this case, unlike that of the coincidence, A's act was the cause of B's death, since each stage is the effect of the preceding stage. Thus, the blow forced the victim against the tree, the effect of this was to make the tree fall and the fall of the tree killed the victim.

---

[1] Above, pp. 67 et seq.

One further criterion in addition to these four must be satisfied if a conjunction of events is to rank as a coincidence and as a limit when the consequences of the action are traced. This further criterion again shows the strength of the influence which the analogy with the case of the simple manipulation of things exerts over thought in causal terms. An abnormal *condition* existing at the time of a human intervention is distinguished both by ordinary thought and, with a striking consistency, by most legal systems[1] from an abnormal event or conjunction of events subsequent to that intervention; the former, unlike the latter, are not ranked as coincidences or 'extraneous' causes when the consequences of the intervention come to be traced. Thus A innocently gives B a tap over the head of a normally quite harmless character, but because B is then suffering from some rare disease the tap has, as we say, 'fatal results'. In this case A has caused B's death though unintentionally. The scope of the principle which thus distinguishes contemporaneous abnormal conditions from subsequent events is unclear; but at least where a human being initiates some physical change in a thing, animal, or person, abnormal physical states of the object affected, existing at the time, are ranked as part of the circumstances in which the cause 'operates'. In the familiar controlling imagery these are part of 'the stage already set' before the 'intervention'.[2]

Judgments about coincidences, though we often agree in making them, depend on two related ways on issues incapable of precise formulation. One of these is patent, the other latent but equally important. Just how unlikely must a conjunction be to rank as a coincidence, and in the light of what knowledge is likelihood to be assessed? The only answer is: 'very unlikely in the light of the knowledge available to ordinary men.' It is of course the indeterminacies of such standards, implicit in causal judgments, that make them inveterately disputable, and call for the exercise of discretion or choice by courts. The second and latent indeterminacy of these judgments depends on the fact that the things or events which they relate do not have pinned to them some uniquely correct description always to be used in assessing likelihood. It is an important pervasive feature of all our empirical judgments that there is a constant possibility of more or less specific description of any event or thing with which they are concerned. The tree might be described not simply as 'a tree' but as a 'rotten tree' or as a 'fir tree' or a 'tree sixty feet tall'. So too its

[1] Footnote omitted.

[2] Footnote omitted.

fall might be described not as a 'fall' but as a fall of a specified distance at a specified velocity. The likelihood of conjunctions framed in these different terms would be differently assessed. The criteria of appropriate description like the standard of likelihood are supplied by consideration of common knowledge. Even if the scientist knew the tree to be rotten and could have predicted its fall with accuracy, this would not change the judgment that its fall at the time when B was struck down within its range was a coincidence; nor would it make the description 'rotten tree' appropriate for the assessment of the chances involved in this judgment. There are other controls over the choice of description derived from the degree of specificity of our interests in the final outcome of the causal process. We are concerned with the fall of an object sufficient to cause 'death' by impact and the precise force or direction which may account for the detail of the wounds is irrelevant here.

Opportunities and Reasons

Opportunities. The discrimination of voluntary interventions as a limit is no longer made when the case, owing to the commonness or appreciable risk of such harmful intervention, can be brought within the scope of the notion of providing an opportunity, known to be commonly exploited for doing harm. Here the limiting principles are different. When A leaves the house unlocked the range of consequences to be attributed to this neglect, as in any other case where precautions are omitted, depends primarily on the way in which such opportunities are commonly exploited. An alternative formulation of this idea is that a subsequent intervention would fall within the scope of consequences if the likelihood of its occurring is one of the reasons for holding A's omission to be negligent.

It is on these lines that we would distinguish between the entry of a thief and of a murderer; the opportunity provided is believed to be sufficiently commonly exploited by thieves to make it usual and often morally or legally obligatory not to provide it. Here, in attributing consequences to prior actions, causal judgments are directly controlled by the notion of the risk created by them. Neglect of such precautions is both unusual and reprehensible. For these reasons it would be hard to separate the two ways in which such neglect deviates from the 'norm'. Despite this, no simple identification can be made of the notion of responsibility with the causal connexion which is a ground for it. This is so because the provision of an opportunity commonly taken by others is ranked as the cause of the outcome independently of the wish to praise or blame. The causal judgment may be made simply to assess a contribution to some outcome. Thus, whether we think well or ill of the use made of railways, we would still claim that the greater mobility of the population in the nineteenth century was a consequence of their introduction.

It is obvious that the question whether any given intervention is a sufficiently common exploitation of the opportunity provided to come within the risk is again a matter on which judgments may differ though they often agree. The courts, and perhaps ordinary thought also, often describe those that are sufficiently common as 'natural' consequences of the neglect. They have in these terms discriminated the entry of a thief from the entry of a man who burnt the house down, and refused to treat the destruction of the house as a 'natural' consequence of the neglect.[1]

We discuss later in Chapter IX the argument that this easily intelligible concept of 'harm within the risk', over-riding as it does the distinctions between voluntary interventions and others, should be used as the general test for determining what subsequent harm should be attributed for legal purposes to prior action. The merits of this proposal to refashion the law along these simple lines are perhaps considerable, yet consequences of actions are in fact often traced both in the law and apart from it in other ways which depend on the discrimination of voluntary interventions from others. We distinguish, after all, as differing though related grounds of responsibility, causing harm by one's own action and providing opportunities for others to do harm, where the guiding analogy with the simple manipulation of things, which underlies causal thought, is less close. When, as in the examples discussed above, we trace consequences through the non-voluntary interventions of others our concern is to show that certain stages of the process have a certain type of connexion with the preceding stages, and not, as when the notion of risk is applied, to show that the ultimate outcome is connected in some general way with the initiating action. Thus, when A's shot makes B start and break a glass it is the causal relationship described by the expression 'made B start' that we have in mind and not the likelihood that on hearing a shot someone may break a glass. Causal connexion may be traced in such cases though the initiating action and the final outcome are not contingencies that commonly go together.

Apart from these conceptual reasons for distinguishing these related grounds for responsibility, it is clear that both in the law, as we show in Chapter VI, and apart from it we constantly treat harm as caused by a person's action though it does not fall 'within the risk'. If, when B broke the glass in the example given above, a splinter flew into C's eye, blinding him, A's action is indeed the cause of C's injury though we may not always blame him for so unusual a consequence.

[1] *Bellows* v. *Worcester Storage Co.* (1937), 297 Mass. 188, 7 N.E. 2d 588.

Reasons. In certain varieties of interpersonal transactions, unlike the case of coercion, the second action is quite voluntary. A may not threaten B but may bribe or advise or persuade him to do something. Here, A does not 'cause' or 'make' B do anything: the strongest words we should use are perhaps that he 'induced' or 'procured' B's act. Yet the law and moral principles alike may treat one person as responsible for the harm which another free agent has done 'in consequence' of the advice or the inducements which the first has offered. In such cases the limits concern the range of those actions done by B which are to rank as the consequence of A's words or deeds. In general this question depends on A's intentions or on the 'plan of action' he puts before B. If A advises or bribes B to break in and steal from an empty house and B does so, he acts in consequence of A's advice or bribe. If he deliberately burns down the house this would not be treated as the consequence of A's bribe or advice, legally or otherwise, though it may in some sense be true that the burning would not have taken place without the advice or bribe. Nice questions may arise, which the courts have to settle, where B diverges from the detail of the plan of action put before him by A.

Fowler V. Harper and Fleming James, Jr., *The Law of Torts* (Boston. Little, Brown and Company, 1956), pp 1108-1121.

§20.1. Introductory. Negligence is not a ground of liability unless it causes injury or damage to some interest which the law recognizes and protects. Moreover, it does not make a defendant liable for any injury or damage that is not a consequence of the negligence. The establishment of the requisite causal connection is therefore an element of a plaintiff's cause of action for negligence, to be pleaded and proven by him. And where a cause of action exists, the question of causal connection will determine the scope of liability — the extent of the injury or damage for which defendant will have to pay. The question then naturally arises what is the kind of causal connection or relationship that the law requires to be proven. Obviously the legal test includes a requirement that the wrongful conduct must be a *cause in fact* of the harm; but if this stood alone the scope of liability would be vast indeed, for "the causes of causes [are] infinite" — "the fatal trespass done by Eve was cause of all our woe." But the law has not stopped there — it has developed further restrictions and limitations. The concept this development has produced is generally called "proximate" or "legal" cause. To be sure this concept is only one of the devices used to limit the fact and the extent of liability for negligence.[1] In the progress of negligence law, however, the concept of proximate cause has been greatly overworked

§20.1. [1] Limitations on the scope of duty and the rule of contributory negligence are, for example, other such devices.

to limit or control both the liability of defendant and the effect of contributory negligence because of many considerations which can be treated in a more meaningful and significant way in connection with other issues, such as that of duty, standard of conduct and the like.[2] "Having no integrated meaning of its own, [the] chameleon quality [of proximate cause] permits it to be substituted for any one of the elements of a negligence case when decision on that element becomes difficult. . . . No other formula . . . so nearly does the work of Aladdin's lamp." [3] The result has been a widely recognized confusion, and as luxuriant a crop of legal literature as is to be had in any branch of tort law.[4] Perhaps recent years have seen a little headway made in dispelling the confusion and taking some of the work load off of this weary concept by separating other strands for analysis and treatment under other more appropriate headings. If this is so, much of the credit belongs to Dean Green.

---

[2] See Green, Rationale of Proximate Cause 78, 122 et seq. (1927); Carpenter, Workable Rules for Determining Proximate Cause, 20 Calif. L. Rev. 229, 246–255 (1932); and Maltbie, C.J., in Kinderavitch v. Palmer, 127 Conn. 85, 89 et seq., 15 A.2d 83, 86 (1940).

[3] Green, Proximate Cause in Texas Negligence Law, 28 Texas L. Rev. 471, 471–472 (1950).

[4] The following is a suggestive but not exhaustive list of literature: Green, Rationale of Proximate Cause (1927); Bohlen, The Probable or the Natural Consequences as the Test of Liability in Negligence, 49 Am. L. Reg. 79, 148 (1901); Bingham, Some Suggestions Concerning "Legal Cause" at Common Law, 9 Colum. L. Rev. 16, 136 (1909); Smith, Legal Cause in Actions of Tort, 25 Harv. L. Rev. 103, 223, 303 (1911); Terry, Proximate Consequences in the Law of Torts, 28 Harv. L. Rev. 10 (1914); Beale, The Proximate Consequences of an Act, 33 Harv. L. Rev. 633 (1920); Pollock, Liability for Consequences, 38 L.Q. Rev. 165 (1922); Levitt, Cause, Legal Cause and Proximate Cause, 21 Mich. L. Rev. 34, 160 (1922); Reese, Negligence and Proximate Cause, 7 Cornell L.Q. 95 (1922); Green, Are Negligence and "Proximate" Cause Determined by the Same Test, 1 Texas L. Rev. 242, 256 (1923); Edgerton, Legal Cause, 72 U. Pa. L. Rev. 211, 343 (1924); McLaughlin, Proximate Cause, 39 Harv. L. Rev. 149 (1925); Burke, Rules of Legal Cause in Negligence Cases, 15 Calif. L. Rev. 1 (1926); Green, Are There Dependable Rules of Causation, 77 U. Pa. L. Rev. 601 (1929); Carpenter, Workable Rules for Determining Proximate Cause, 20 Calif. L. Rev. 229, 396, 471 (1932); Peaslee, Multiple Causation and Damage, 47 Harv. L. Rev. 1127 (1934); Carpenter, Concurrent Causation, 83 U. Pa. L. Rev. 941 (1935); Prosser, The Minnesota Court on Proximate Cause, 21 Minn. L. Rev. 19 (1936); Campbell, Duty, Fault and Legal Cause, [1938] Wis. L. Rev. 402; Gregory, Proximate Cause in Negligence — A Retreat from Rationalization, 6 U. Chi. L. Rev. 36 (1938); Morris, On the Teaching of Legal Cause, 39 Colum. L. Rev. 1087 (1939); Carpenter, Proximate Cause, 14 So. Calif. L. Rev. 1, 115 (1940), 15 So. Calif. L. Rev. 187, 304, 427 (1941), 16 So. Calif. L. Rev. 1, 61, 275 (1942); Morris, Proximate Cause in Minnesota, 34 Minn. L. Rev. 185 (1950); Green, Proximate Cause in Texas Negligence Law, 28 Texas L. Rev. 471, 621, 755 (1950); Morris, Duty, Negligence and Causation, 101 U. Pa. L. Rev. 189 (1952); Green, Proximate Cause, 17 Texas B.J. 161 (1954).

These writings represent nearly every conceivable point of view. For reasons that the present authors find elusive, Morris goes so far as to disparage the requirement of cause in fact.

§20.2. **Causal relation or cause in fact.** Through all the diverse theories of proximate cause runs a common thread; all agree that defendant's wrongful conduct must be a cause in fact of plaintiff's injury before there is liability. This notion is not a metaphysical one but an ordinary, matter-of-fact inquiry into the existence or nonexistence of a causal relation as laymen would view it. Clearly this is not a quest for a *sole* cause. Probably it cannot be said of any event that it has a single causal antecedent; usually there are many. For the purpose of the present inquiry it is enough that defendant's negligence be *a* cause in fact of the harm.

A rough working test of this relation, valid for most cases and enjoying wide currency, is the "but for" or *sine qua non* test: defendant's negligence is a cause in fact of an injury where the injury would not have occurred *but for* defendant's negligent conduct.[1] It is probably safe to say that wherever this test is met, the cause in fact relation[2] does exist. But the test is not universally valid as a test of exclusion of causes in fact. There is one situation where it will not work. If two independent causes concur to produce a result which either of them alone would have produced, each is a cause in fact of that result though it would have happened without *either* cause (but not if *both* causes were absent).[3] Thus if two fires from separate sources combine as they reach plaintiff's property, and consume it, each is a cause though it be assumed or shown that the property would have been consumed as completely by either fire alone.

So far as the *substantive* law of torts goes, the cause in fact aspect of the requirement of legal cause gives little trouble, and it has been called a simple one.[4] Yet there are problems here which often beset the trial court and practitioner and prove fatal to many a case. The most serious of these problems is that of sufficiency of proof where the ascertainable facts are meager, or where they present complicated questions of science, medicine, engineering, or the like. Another — usually less serious — problem is in finding appropriate language for the charge.

---

§20.2. [1] Carpenter, Workable Rules for Determining Proximate Cause, 20 Calif. L. Rev. 229, 235–236, 396–397 (1932); Smith, Legal Cause in Actions of Tort, 25 Harv. L. Rev. 103, 108–110 (1911).

[2] But not necessarily the relationship of legal or proximate cause. It has been said that the "but for" test is valid as a test of exclusion (but not of inclusion) for *proximate* cause (except in the case of the concurring efficient causes given in the text). Smith, Legal Cause in Actions of Tort, 25 Harv. L. Rev. 103, 109 (1911). A moment's reflection will reveal that this statement and the one in the text (which is concerned with cause *in fact* only) are not at all inconsistent. Carpenter, Workable Rules for Determining Proximate Cause, 20 Calif. L. Rev. 229, 396 (1932). remarks, obliquely, that the "but for" rule may be used as a test for *exclusion* with respect to proximate causes and of *inclusion* of causes in fact.

[3] See Prosser, Torts 220 (2d ed. 1955).

[4] Green, Rationale of Proximate Cause 136 (1927); Green, Proximate Cause in Texas Negligence Law, 28 Texas L. Rev. 471, 474, 476 (1950). Of course, if unwarrantably confused with other aspects of a negligence case, the "causation" issue thus created may appear extremely complex.

The problem of proof here is essentially like that presented by the problem of proving negligence circumstantially, and the two often overlap. Thus in *res ipsa loquitur* cases the requirement of defendant's exclusive control of the injuring agency really calls for nothing more than proof tending to eliminate other possible causes of the occurrence, so as to indicate that the negligence of which that occurrence speaks is probably that of defendant.[5] But where negligence is pinned on the defendant by direct or other sufficient proof there may still remain a serious problem of proving the causal relation. This is not always so — perhaps in the bulk of accident cases the real issues turn on credibility of witnesses and the evaluation of conduct.[6] But there is a substantial number where causal relation rests on inference from circumstantial evidence, and in them is found the same test of delusive exactness that is generally used to gauge the legitimacy of the basis for an inference from circumstantial evidence. "Where," the New York court has said, "the facts proven show that there are several possible causes of an injury, for one or more of which the defendant was not responsible, and it is just as reasonable that the injury was the result of one cause or the other, plaintiff cannot have a recovery since he has failed to prove that the negligence of the defendant caused the injury."[7] But here, as elsewhere,

---

[5] See §19.7 *supra;* James, Proof of the Breach in Negligence Cases, 37 Va. L. Rev. 179, 204 et seq. (1951). The test of legitimacy of inferences from circumstantial evidence is discussed in §19.4 *supra.*

[6] Thus, where there are eyewitnesses to an accident, their testimony will generally suffice to establish defendant's breach of duty and direct causation of plaintiff's injuries.

[7] Ingersoll v. Liberty Bank of Buffalo, 278 N.Y. 1, 7, 14 N.E.2d 828, 829 (1938). See also Alling v. Northwestern Bell Tel. Co., 156 Minn. 60, 63, 194 N.W. 313, 314 (1923) ("The burden is on the plaintiff to show that it is more probable that the harm resulted in consequence of something for which the defendant was responsible than in consequence of something for which he was not"), and Ramberg v. Morgan, 209 Iowa 474, 487, 218 N.W. 492, 498 (1928) ("Where the proof discloses that a given result may have occurred by reason of more than one proximate cause, and the jury can do no more than guess or speculate as to which was, in fact, the efficient cause, the submission of such choice to the jury has been consistently condemned by this court and by other courts"; quoted with approval in Frye v. City of Detroit, 256 Mich. 466, 470, 239 N.W. 886, 887 (1932).

Cf. the more liberal statement of the test found in Lunde v. Cudahy Packing Co., 139 Iowa 688, 701, 117 N.W. 1063, 1068 (1908) and quoted with approval in Bock v. Fellman Dry Goods Co., 212 S.W. 635, 637 (Tex. Civ. App. 1919) ("A cause being this test permits wide latitude where facts are meager and where it is pure matter of guesswork where the greater probabilities lie. A man is found dead at the foot of an unlighted apartment hallway,[8] or at the bottom of an elevator well one of the entrances to which was left unguarded,[9] or at a railroad crossing after the passing of a train that failed to whistle.[10] Death occurred at the unsafe place, or at the hands of the negligently operated agency, but did it occur because of the defect or the negligence? Would the precautions demanded by reasonable care have averted the accident? Obviously no one knows — moreover no one even knows what the probabilities are, and the courts are free to choose either a postulate which will let the jury draw an inference of negligence, or its opposite.

On the meager kind of showing described, therefore, it is not surprising to find variant results.[11] Courts which take a strict attitude stress the fallacy of reasoning post hoc ergo propter hoc,[12] and the likelihood of explanations involving no causal relation: "It is a matter of common knowledge that people sometimes fall

---

shown which might produce an accident, and it further appearing that an accident of that particular character did occur, it is a warrantable inference, in the absence of showing of other cause, that the one known was the operative agency in bringing about such results").

[8] Ingersoll v. Liberty Bank of Buffalo, 278 N.Y. 1, 14 N.E.2d 828 (1938); Wolf v. Kaufmann, 227 App. Div. 281, 237 N.Y. Supp. 550 (1929); cf. Bornstein v. Faden, 149 App. Div. 37, 133 N.Y. Supp. 608 (1912), *aff'd,* 208 N.Y. 605, 102 N.E. 1099 (1913); Reynolds v. Texas & P. Ry. Co., 37 La. Ann. 694 (1885); Paine v. Gamble Stores, 202 Minn. 462, 279 N.W. 257 (1938).

[9] Hopkinson v. Knapp & Spalding Co., 92 Iowa 328, 60 N.W. 653 (1894); Bock v. Fellman Dry Goods Co., 212 S.W. 635 (Tex. Civ. App. 1919); McCowat-Mercer Printing Co. v. Taylor, 115 F.2d 868 (6th Cir. 1940).

[10] Puckhaber v. So. Pac. R. Co., 132 Cal. 363, 64 Pac. 480 (1901); Burlington-Rock Island R. Co. v. Ellison, 140 Tex. 353, 167 S.W.2d 723 (1943). Cf. Woodworth v. N.Y. Central R. Co., 149 Ohio St. 543, 80 N.E.2d 142 (1948), soundly criticized in a Note, 18 U. Cin. L. Rev. 371 (1949).

[11] On nearly identical facts even the same judge may come to contradictory conclusions at different times. Compare Wolf v. Kaufmann, 227 App. Div. 281, 237 N.Y. Supp. 550 (1929) (denying recovery for death of plaintiff's intestate found unconscious at foot of stairway unlighted in violation of statute) with Ingersoll v. Liberty Bank of Buffalo, 278 N.Y. 1, 14 N.E.2d 828 (1938) (directly opposed). Both opinions were by Finch, J.

[12] See, e.g., Springer v. Security Nat. Bank etc. Co., 175 S.W.2d 797 (1943) (connection between alleged negligence of broker and loss to customer not established); Payne v. Chandler, 41 Ga. App. 385, 153 S.E. 96 (1930) (evidence that plaintiff suffered pain in heart and other ailments after dentist caused her to swallow unidentified, unpleasant-tasting liquid is insufficient to authorize inference of causal relation); Kramer Service v. Wilkins, 184 Miss. 483, 186 So. 625 (1939) (connection between traumatic injury to hand and cancer of skin at point of injury not established).

But cf. Reynolds v. Texas & P. Ry. Co., 37 La. 694 (1885), in which the court recognized the "distinction between *post hoc* and *propter hoc,*" but found for the plaintiff anyway because "the whole tendency of the evidence connects the accident with the negligence."

downstairs in broad daylight";[13] the pedestrian who failed to hear the train might well not have heard the whistle, perhaps he fainted on the tracks,[14] and so on.[15] Courts which take a sympathetic view point out that defendant's negligence greatly multiplied the chances of an accident of just the kind plaintiff claims happened, that the proven facts are entirely consistent with its having happened that way, and that the evidence does not strongly suggest an alternative explanation.[16]

In most cases the facts are not as meager as those described above and the additional evidence generally reduces the area of uncertainty and enables the court to make a much more surefooted judgment on the matter of probabilities, though even here there is often room for difference of opinion. On the whole, courts have tended to view with liberality the legitimacy of the inference of causal relation in these cases. Thus where it appears *how* an accident happened and also that the victim *might* have saved himself by taking advantage of a precaution which it has been shown defendant negligently failed to afford, courts have generally let a jury find the failure caused the harm, though it is often a pretty speculative matter whether the precaution would in fact have saved the victim.[17] Of course if the possibility of

[13] Dissenting opinion by McLaughlin, J., in Bornstein v. Faden, 149 App. Div. 37, 42, 133 N.Y. Supp. 608, 611 (1912), *aff'd*, 208 N.Y. 605, 102 N.E. 1099 (1913). See also Wolf. v. Kaufmann, 227 App. Div. 281, 237 N.Y. Supp. 550 (1929).

[14] Puckhaber v. So. Pac. R. Co., 132 Cal. 363, 64 Pac. 480 (1901).

[15] For example, the unsympathetic rejection of inferences as to the cause of the airplane crash in which the plaintiff's decedent was killed in Morrison v. LeTourneau Co. of Georgia, 138 F.2d 339 (5th Cir. 1943), criticized in Note, 38 Ill. L. Rev. 326 (1944).

[16] See Reynolds v. Texas & P. Ry. Co., 37 La. Ann. 694, 698 (1885) ("where the negligence of the defendant greatly multiplies the chances of accident to the plaintiff, and is of a character naturally leading to its occurrence, the mere possibility that it might have happened without the negligence is not sufficient to break the chain of cause and effect between the negligence and the injury"); Lunde v. Cudahy Packing Co., 139 Iowa 688, 117 N.W. 1063 (1908) (statement quoted in note 7 *supra*); Emery v. Tilo Roofing Co., 89 N.H. 165, 195 Atl. 409 (1937); Gates v. B. & M.R. Co., 255 Mass. 297, 151 N.E. 320 (1926); Burlington-Rock Island R. Co. v. Ellison, 140 Tex. 353, 167 S.W.2d 723 (1943); Texas & N.O.R. Co. v. Ewing, 46 S.W.2d 398 (Tex. Civ. App. 1931); Bornstein v. Faden, 149 App. Div. 37, 133 N.Y. Supp. 608 (1912), *aff'd*, 208 N.Y. 605, 102 N.E. 1099 (1913); Cornbrooks v. Terminal Barber Shops, Inc., 282 N.Y. 217, 26 N.E.2d 25 (1940); Note, 48 Mich. L. Rev. 536 (1950). Cf. note 38 *infra*.

[17] See 2 Restatement of Torts §432, Comment *c*. In Zinnel v. U. S. Shipping Board, 10 F.2d 47, 49 (2d Cir. 1925), the issue was whether a guard rope, the absence of which constituted the alleged negligence of the defendant, would have prevented the plaintiff's intestate from being washed overboard. L. Hand, J., said, in reversing a judgment dismissing the complaint: "About that we agree no certain conclusion was possible. Nobody could, in the nature of things, be sure that the intestate would have seized the rope, or, if he had not, that it would have stopped his body. But we are not . . . justified, where certainty is impossible, in insisting upon it. We cannot say that there was no likelihood that a rope three feet above

salvation is too clearly foreclosed by the facts of a given case, the inference may not be drawn.[18] Grade crossing collisions furnish familiar examples. Where the victim's automobile is driven onto the crossing immediately in front of an obvious and very audible train the jury still may find that the driver's oblivious attention would have been arrested in time if the whistle had been blown.[19]

---

the deck . . . would not have saved the seaman. . . . Considering that such lines were run for the express purpose, among others, of protecting seamen, we think it a question about which reasonable men might at least differ whether the intestate would not have been saved had it been there." See also Kirincich v. Standard Dredging Co., 112 F.2d 163 (3d Cir. 1940), wherein a seaman who could not swim fell overboard and was tossed a one-inch heaving line which landed two feet from where he was struggling; the court found failure to heave an available ring buoy to be the cause in fact of the drowning. ". . . we think he might (the appropriate grammatical mood) have saved himself through the help of something he could more easily grab." 112 F.2d at 164. Cf. Harris v. Pennsylvania R. Co., 50 F.2d 866 (4th Cir. 1931) (semble, but Harris was a good swimmer and therefore there was even greater probability of his reaching a ring buoy or heaving line, both of which were available, had defendant's servants thrown one).

[18] N.Y. Cent. R. Co. v. Grimstad, 264 Fed. 334 (2d Cir. 1920), contrasts nicely with the cases cited in note 17 *supra* and illustrates the point of difference made in the text. The man overboard could not swim, but was seen by his wife while he was still struggling. She rushed back into the cabin for a line, but when she returned he had disappeared. The negligence alleged was the failure to equip the barge with a life buoy, but the court found any causal connection between this failure and the drowning to be "pure conjecture and speculation." There was nothing to show that, if there had been a life buoy on board, the wife would have got it sooner than she got the small line. Ford v. Trident Fisheries Co., 232 Mass. 400, 122 N.E. 399 (1919) is similar (negligence with regard to readiness of lifeboat immaterial as man disappeared immediately upon falling overboard, and therefore could not have been saved even if defendant had not been negligent).

It will be recognized that these cases illustrate the converse of the "but for" rule: If the accident would have happened without defendant's negligent act, then such is not the cause of it. See Russo v. Aucoin, 7 So. 2d 744 (La. App. 1942) (negligence of D_1 in pulling away from edge of road without signaling held not to be cause of accident in which D_2, going southward, as was D_1, came speeding around blind curve, saw D_1 in his path, but being unable to stop by reason of defective brakes, grazed D_1's car and plowed into the plaintiff's truck, which was going northward; even if D_1 had signaled, the accident would have happened the same way); Fleeherty v. Sullivan, 129 S.W.2d 926 (Mo. App. 1939) (failure to warn by sounding horn not proximate cause of injury where person to be warned and complaining of the failure was aware of the impending danger); Stacy v. Knickerbocker Ice Co., 84 Wis. 614, 54 N.W. 1091 (1893) (failure to indicate location of, or to fence in, hole in ice not the cause of the drowning of runaway horses since the horses could not have been halted or saved had all these precautions been taken); City of Piqua v. Morris, 98 Ohio St. 42, 120 N.E. 300, 7 A.L.R. 129 (1918) (negligence of city in construction or maintenance of spillway not the cause of overflow of pond to plaintiff's damage, where extraordinary rainfall would have caused pond to overflow its embankment in any event); West Texas Utilities Co. v. Harris, 231 S.W.2d 558 (Tex. Civ. App. 1950) (failure to maintain high voltage wire at statutory height not cause of decedent's electrocution where pipe handled by decedent would have touched wire even if it had been placed at statutory height). But see Note, 29 Texas L. Rev. 691 (1951).

[19] Davis v. Boggs, 22 Ariz. 497, 199 Pac. 116 (1921); Johnston v. Southern Pac. Co., 105 Cal. App. 340, 288 Pac. 81 (1940). Cf. Faust v. Philadelphia & R. Ry. Co., 191 Pa. 420, 43 Atl. 329 (1899); Gregory v. Seaboard Air Line Ry. Co., 142 Va. 750, 128 S.E. 272 (1925).

But where the automobile is driven into the sixtieth car of a freight train at an open and unobstructed crossing in the night time, the absence of an unlighted warning sign could scarcely be the cause of the accident. "[I]f travelers on the highway could not see the moving cars ahead of [them], they could not see the sign."[20]

A recent case has perhaps introduced a new technique for dealing with the "equal probability" problem, at least in a limited type of situations. In *Summers v. Tice,*[21] two members of a hunting party fired simultaneously at game. A pellet of shot entered the eye of a third member of the party whom the others should have known to be dangerously near their line of fire. There was no way to tell from whose gun the pellet came. In a suit against both, the California court held them both liable to plaintiff. Yet only one defendant caused the damage and in the case against each it is obvious that plaintiff could not prove this element by the *greater probability.* The result actually does no more violence to the equal probability rule than many of the instances just cited — it simply conflicts with that rule more dramatically. The court frankly faced this problem, pointed out that ordinarily defendants would be in a far better position than plaintiff to offer evidence as to cause, and concluded, "When we consider the relative position of the parties and the results that would flow if plaintiff was required to pin the injury on one of the defendants only, a requirement that the burden of proof on that subject be shifted to defendants becomes manifest."[22] This is like the technique used by the same court in the *res ipsa loquitur* case of *Ybarra v. Spangard*[23] but has even greater claim to acceptance

---

Cf. cases in which the approach of a train was unknown to the plaintiff until both he and it were at the crossing, the train having negligently failed to signal until it was almost upon the crossing, and the noise of the belated signal instead of alerting and saving the plaintiff caused his injury. Roberts v. Chicago & N.W.R. Co., 35 Wis. 679 (1874) (plaintiff's horses frightened by train's signal); Walling v. Central R. Co. of N.J., 82 N.J.L. 506, 81 Atl. 987 (1911) (semble); Louisiana & Ark. Ry. Co. v. Nix, 94 Ark. 270, 126 S.W. 1076 (1910) (semble).

Of course, where the engine was not visible or audible — e.g., if it is diesel-powered or a wind of great intensity is blowing — the jury will be allowed to find the failure of the train to sound a warning to be the proximate cause of a death occurring at a crossing. See Burlington-Rock Island R. Co. v. Ellison, 140 Tex. 353, 167 S.W.2d 723 (1943).

[20] Texas & N.O.R.R. v. Compton, 135 Tex. 7, 11, 136 S.W.2d 1113, 1115 (1940), favorably commented upon by Green, Proximate Cause in Texas Negligence Law, 28 Texas L. Rev. 471, 477 (1950).

[21] 33 Cal. 2d 80, 199 P.2d 1 (1948).

[22] 33 Cal. 2d at 86, 199 P.2d at 4.

[23] 25 Cal. 2d 486, 154 P.2d 687 (1944), 93 Cal. App. 43, 208 P.2d 445 (1949). In this case, discussed in §19.7 *supra,* and James, Proof of the Breach in Negligence Cases, 37 Va. L. Rev. 179, 207 et seq. (1951), the plaintiff sued the four doctors and two nurses who attended him while he was undergoing appendectomy for damages for a traumatic injury to his shoulder suffered while he was unconscious. The

here where both defendants have been shown negligent.[24] The alternative would turn the innocent victim away without redress and exonerate two admitted wrongdoers, one of whom clearly caused the harm. Both the fault principle and the objective of compensation are flagrantly violated by any such result, and it would be a sterile and rigid procedure indeed that would force it in the teeth of all the policies of the substantive law.[25] If the reasoning of the California court is accepted, interesting questions will come up as to its limits: Will it apply to cases where one or more of the wrongdoers is not before the court as a defendant? Will it cover the case where each of several defendants has caused some harm but it cannot be told how much?[26]

In this matter of proof of causal relations, as in other issues, qualified opinion evidence may be received in appropriate cases.[27] In connection with it two questions of sufficiency may arise. If in any given case the question of causal relation is so esoteric that lay minds cannot form any intelligent judgment about it without

---

court said that under the circumstances "all those defendants who had any control over his body or the instrumentalities which might have caused the injury may properly be called upon to meet the inference of negligence by giving an explanation of their conduct." 25 Cal. 2d at 494, 154 P.2d at 691.

[24] Seavey says he knows of no other case "in which it has been held that the plaintiff has sustained the burden of proof where there was no evidence that it was more likely than not that a defendant caused the plaintiff's harm." Seavey, Res Ipsa Loquitur, Tabula in Naufragio, 63 Harv. L. Rev. 643, 648 (1950). On the ground stated in the text, however, he feels that the Summers case can be more readily supported than the Ybarra case. Ibid. Any such distinction rests, of course, on accepting the importance of fault as a criterion of liability. The reasoning in the Summers case would probably have been approved by such authorities as Wigmore and Carpenter. See Note, 17 Ill. L. Rev. 458 (1923); Carpenter, Workable Rules for Determining Proximate Cause, 20 Calif. L. Rev. 229, 396, 406 (1932).

Other cases have reached a like result where the plaintiff was not a member of the party, on a theory of concerted action or joint enterprise. Benson v. Ross, 143 Mich. 452, 106 N.W. 1120 (1906); Oliver v. Miles, 144 Miss. 852, 110 So. 666 (1927); cf. State v. Newberg, 129 Ore. 564, 278 Pac. 568 (1929); Regina v. Salmon, 6 Q.B.D. 79 (1880).

[25] It should never be forgotten that procedure should be the handmaiden, not the mistress, of justice. This does not mean, of course, that the fundamental guarantees of procedural due process are not vitally important. They are. But an observance of them will always mean that rights and liabilities under the substantive law will in some cases be unenforceable. And that is a *cost, not an advantage,* of these guarantees. Whenever a procedural rule will consistently work to subvert substantive justice, it should be scrutinized to see whether any fundamental guarantee requires the rule to persist in such a form. Here clearly it does not. Even granting sacredness to the fault principle, there are many instances where procedural burdens have been shifted by statute or judicial ruling on the basis of considerations like those present here. Res ipsa loquitur and the presumption against bailees furnish ready examples. For a narrow view of the Summers case, however, see Note, 37 Geo. L.J. 627 (1948).

[26] See the opinion of the court in the Ybarra case, 25 Cal. 2d 486, 494, 154 P.2d 687, 691 (1944).

[27] See note 29 *infra* for some recent cases in which it has played a significant role in the cause issue.

expert aid, opinion evidence from qualified witnesses may be required as it sometimes is on the question of appropriate standards of conduct.[28] Probably no more crystallized rule than this can be formulated. The courts have proceeded on a case to case basis, and there are few decisions denying recovery which have laid stress on the want of expert evidence as to the cause in fact of the accident.[29] It should be noted that where the doctrine of *res ipsa loquitur* is applied on the basis of judicial notice, expert evidence is dispensed with upon this issue as upon any other.[30] Expert evidence is often required to establish the causal connection between the accident and some item of physical or mental injury[31] unless the connection is a kind which would be obvious to laymen, such as a broken leg from being struck by an automobile.

Where opinion evidence is relied on to establish the causal link a question of sufficiency may also arise in deciding whether the opinion has been expressed in terms definite enough to support the needed conclusion. If a qualified witness gives it as his opinion that the causal relation *did exist* or *probably did exist*, there is not much question about the legal sufficiency of the testimony.[32] But cautious or hostile witnesses are often unwilling to go so far, and there are serious questions and variant rulings about the sufficiency of such expressions as "may have caused," "possibly caused," and the like.[33] Of course other factors in the case may lend support to the inference. Thus in a miscarriage case the Massachusetts court has held such an expression sufficient when "taken in connection with the plaintiff's testimony that [her] health was good before the accident" though otherwise it would leave "the issue trembling in the balance."[34]

---

[28] See Morris, The Role of Expert Testimony in the Trial of Negligence Issues, 26 Texas L. Rev. 1, 10 et seq. (1947).

[29] Two such cases are Blarjeske v. Thompson's Restaurant Co., 325 Ill. App. 189, 59 N.E.2d 320 (1945) (expert was not asked his opinion as to whether condition of meat caused plaintiff's illness; *held*, question of proximate cause left in realm of speculation, surmise and conjecture even though nearly all other evidence pointed very strongly to a causal relation), and Goodwin v. Misticos, 207 Miss. 361, 42 So. 2d 397, 402 (1949). ("The fact whether the matter vomited up by Mr. Goodwin contained corned beef that was infected by poisonous bacteria was also one capable of direct and demonstrative proof by a chemical analysis"). There are several recent cases in which such evidence was offered where the court might well have found the lack of it fatal. See Foley v. Pittsburgh-Des Moines Co., 363 Pa. 1, 68 A.2d 617 (1949) (expert traced explosion of liquefied natural gas to negligence of manufacturer of tank in which it was stored), and National Lead Co. v. Schuft, 176 F.2d 610 (8th Cir. 1949) (experts ascribed puffs of flame and explosions to large amounts of floating humus dust in the air). Cf. Western Tel. Corp. v. McCann, 128 Tex. 582, 99 S.W.2d 895 (1937) (defendant's experts successfully disproved any causal connection between a stroke of lightning coupled with the admitted negligence of the defendant in the management of its wires and the death of plaintiff's wife by electrocution).

In Comeau v. Beck, 319 Mass. 17, 64 N.E.2d 436 (1946), it was the testimony of the *defendant's* medical expert to the effect that a miscarriage might be produced by "'some injury, [or by] the striking of the abdomen,' [which] was enough to support a finding that the plaintiff's miscarriage was causally related to the accident."

[30] See §19.5 *supra;* James, Proof of the Breach in Negligence Cases, 37 Va. L. Rev. 179, 197 n.55, 201, 202 (1951). Cf. Judson v. Giant Powder Co., 107 Cal. 549, 40 Pac. 1020, 29 L.R.A. 718, 48 Am. St. Rep. 146 (1893), and the discussion of this point in Morris, Res Ipsa Loquitur in Texas, 26 Texas L. Rev. 257, 261–262 (1948).

[31] See cases cited in note 33 *infra.*

[32] Tullgren v. Amoskeag Mfg. Co., 82 N.H. 268, 133 Atl. 4 (1926); National Lead Co. v. Schuft, 176 F.2d 610 (8th Cir. 1949); Averdieck v. Barris, 87 Cal. App. 626, 262

A distinction should be emphasized at this point. We are concerned here with the *sufficiency, not the weight* of the evidence, and on familiar principles plaintiff is entitled to have sufficiency tested on the basis of the evidence most favorable to him.[35] It is

---

Pac. 423 (1927); Shepard v. Carnation Milk Co., 220 Iowa 466, 262 N.W. 110 (1935).

In many cases each side has at least one expert witness who testifies concerning the cause of the injury in definite, unequivocal language. The experts thus diametrically opposed, a jury verdict for either party will be upheld. See Harris v. Wood, 214 Minn. 492, 8 N.W.2d 818 (1943) (strong expert testimony to support two alternative theories as to death of decedent, only one of which involved defendant's negligence; jury verdict for plaintiff upheld), discussed in Morris, Proximate Cause in Minnesota, 34 Minn. L. Rev. 185, 187 (1950).

[33] "Possibility" of causal relation insufficient: Mutual Ben. Health & Accident Assn. v. Johnson, 186 So. 297 (Miss. 1939); Wood v. W. E. Joyce Co., 228 App. Div. 729, 239 N.Y. Supp. 110 (1930); Halnan v. New England Tel. & Tel. Co., 296 Mass. 219, 5 N.E.2d 209 (1936); Gerber v. Wloszczynski, 188 Wis. 304, 206 N.W. 206 (1925). In Western Truck Lines v. Berry, 53 Ariz. 216, 87 P.2d 484, 486 (1939) testimony that "one might . . . say that an accident . . . may or may not have anything to do with it" was held "entirely too indefinite and uncertain" to establish the requisite causal relation. A similar result was reached in Di Fazio v. J. G. Brill Co., 133 Pa. Super. 576, 3 A.2d 216 (1938), the court stating that expert testimony that a causal connection was "extremely possible and likely" would be insufficient.

Of course if the cause and result relationship of the type which the plaintiff is urging as the ground of the defendant liability occurs in only one out of one hundred cases there is insufficient evidence of it to go to the jury in the individual case at bar. Kramer Service v. Wilkens, 184 Miss. 483, 186 So. 625 (1939). But a similar result was reached in Howley v. Kantor, 105 Vt. 128, 163 Atl. 628 (1933), even though the percentages greatly favored the plaintiff (eighty per cent of cases involved the more serious type of tumor for which she was asking damages) on the grounds that the expert testimony was not directed to the likelihood of the plaintiff's particular tumor being of the more serious type.

On the other hand in Cornbrooks v. Terminal Barber Shops, Inc., 282 N.Y. 217, 26 N.E.2d 25 (1940), expert evidence that a barber's electric vibrator "would have been a competent producing cause" of a detachment of the retina (together with absence of proof of other probable cause) was held to satisfy the requirement as to probability. The Appellate Division had found this evidence insufficient to establish the causal connection. 255 App. Div. 522, 8 N.Y.S.2d 171 (1938). The case of Wood v. W. E. Joyce Co., *supra*, also involved a claim of traumatic detachment of the retina, with the elimination of other probable causes.

In Dunham v. Village of Canisteo, 303 N.Y. 498, 104 N.E.2d 872 (1952), a doctor's testimony that "it was a reasonable medical certainty that the delay of eighteen hours [in providing medical aid] contributed to Mr. Dunham's death" was held enough to establish a causal relation, although the doctor also admitted that the same result might have happened if there had been prompt medical care.

[34] Comeau v. Beck, 319 Mass. 17, 20, 64 N.E.2d 436, 437 (1946); cf. De Filippo's Case, 284 Mass. 531, 534, 188 N.E. 245, 247 (1933).

[35] 9 Wigmore, Evidence §§2494, 2495 (3d ed. 1940).

only to that version of the evidence therefore that the equal probability test (of the legitimacy of an inference) may properly be applied. It is the jury's function, not the court's, to say whether plaintiff's version of the case is sustained by the greater weight of the evidence.[36] A case therefore in which the evidence most favorable to plaintiff (including the opinion of qualified experts) affords the basis for an inference of cause is not defeated as matter of law by defendant's evidence of a different version of facts or a different theory of cause in fact.[37] Where the evidence stands in that posture, the question of what probably happened in the case is for the jury even though the court might be inclined to accept the defendant's version.[38]

There is another type of question that presents problems of proof of fact, which is sometimes treated under the heading of cause.[39] This is the question found in negligent entrustment and unlicensed operator cases, namely, whether, in the former, the fact of the driver's incompetence can be used as evidence that the accident occurred as a result of this incompetence, and, in the latter cases, whether the lack of license is evidence that the driver was unskillful in a way which contributed to the accident. Defendant entrusts his automobile to one he should know is incompetent and the latter injures plaintiff, a lawful user of the highway. Under these circumstances the owner's act is clearly a breach of duty towards plaintiff. But there remains a further question before liability may be imposed. This may be put in terms of causal relation between the negligence and the harm or in terms of whether the injury arose out of the risk which made defendant's entrustment negligent. In either event the factual inquiry is the same, did the injury occur *because* the driver (borrower) was incompetent? Neither analysis will permit recovery if the driver was acting with all due care on this occasion. The result is that the negligence of the driver must be proven (just as in a case of vicarious liability, but for different reasons). The only question is whether proof of his incompetence which was introduced to show the owner's negligence may also be used as evidence of negligent operation in the particular accident in the case. The courts have said not, since on grounds of doubtful validity the evidence would have been incompetent if offered for the latter purpose.[40]

---

[36] Id. §2551.

[37] Unless, in any given case, the defendant's evidence is so overwhelming that under general principles reasonable minds could not reject it and accept the plaintiff's.

[38] The actual decisions recognize this. National Lead Co. v. Schuft, 176 F.2d 610 (8th Cir. 1949); Sears Roebuck & Co. v. Scroggins, 140 F.2d 718 (8th Cir. 1944); Turner v. Minneapolis St. Ry. Co., 140 Minn. 248, 167 N.W. 1041 (1918); Tullgren v. Amoskeag Mfg. Co., 82 N.H. 268, 133 Atl. 4, 46 A.L.R. 380 (1926); Foley v. Pittsburgh-Des Moines Co., 363 Pa. 1, 68 A.2d 517 (1949). But the reasoning in these cases sometimes fails to describe with accuracy the basis for holding inapplicable the rule that a jury will not be allowed to guess between two equally probable causes. Thus in the Tullgren case the court says this equal probability rule "has no application unless the existence of a sufficient cause or causes for the injury aside from the negligence charged, is conceded or conclusively proved." 133 Atl. at 7. This statement overlooks the fact that it is the weakness of *plaintiff's case* that usually brings the restrictive rule into play and seems to suggest instead that a plaintiff can get to the jury whenever *defendant's explanation* falls short of certainty. Unfortunately, perhaps, the broad literal implications of this statement do not seem to be borne out by the cases. See, however, Carpenter, Workable Rules for Determining Proximate Cause, 20 Calif. L. Rev. 396, 399 (1932). Possibly the court in the Tullgren case only meant to say that "where a cause is present which might produce a condition that has occurred, and there is no other adequate cause proved, the matter is not within the equipoise rule, and it may, under such conditions be found that the only thing suggested by the evidence which could cause an effect found present, did cause it." Turner v. Hartford F. Ins. Co., 185 Iowa 1363, 1377, 172 N.W. 166, 172 (1919). As we have seen, nearly all the cases which employ even this relaxation of the equal probability formula could with equal logic have been decided for the defendant because plaintiff's proof of cause rested on conjecture. For examples, see cases cited in notes 16, 17 and 19 *supra*.

[39] We shall deal with the problem of proof here, and leave for later discussion the problem of whether the question should be assimilated to the issue of cause or the issue of duty. See §20.5, notes 15–17 *infra*.

Where the car is entrusted to an unlicensed driver, or where defendant himself drives it without a license, a similar question is presented (if the licensing statute is thought to have safety in mind and not exclusively revenue, or some other unrelated purpose), namely, did the harm result from the want of a license, or (more accurately) because of the risk of unskilled driving which the licensing statute sought to prevent? If the car was driven with all the skill and care the law requires, then the harm did not proceed from such a risk; if the car was negligently driven, then it did. The further question then arises: may the want of license be considered as *some* evidence that unskillfulness contributed to the accident? Some courts say no.[41] But this effectively deprives the victim of any benefit, in a civil case, of a statute concededly passed in part for his protection. Nor does the ruling seem theoretically sound. The legislature has decided that the general safety requires the activity in question to be limited to those who can demonstrate at least a minimum of special skill. It is fair to assume that the unlicensed as a class are far less likely than those licensed to have the skill for which a license is required. And it is certainly true that when the unskilled attempt what it takes skill to do, some of the intangible factors that go to make up lack of skill are far more likely than not to have contributed to any mishap that occurs. Frequently they do this in ways that are hard to prove; moreover, what evidence there is in the matter is likely to be in the defendant's hands. Both probability and policy, therefore, call for the rule that breach of a licensing statute, if it is negligence to the plaintiff at all, should be *prima facie* evidence that it is the "proximate cause" of any injury that ensues.[42]

---

[40] In nearly all jurisdictions the character or propensity for carelessness of a person charged with a negligent act may not be put in evidence even though it might throw light "on the probability of his having acted carelessly on the occasion in question." See 1 Wigmore, Evidence §65 (3d ed. 1940). This matter is discussed in James and Dickinson, Accident Proneness and Accident Law, 63 Harv. L. Rev. 769, 791 (1950).

[41] Mahowald v. Beckrich, 212 Minn. 78, 2 N.W.2d 569 (1942), lists several of the jurisdictions in which both questions in the text have been answered in the negative. See also Morrison v. Le Tourneau, 138 F.2d 339 (5th Cir. 1943): "There is argument that the possession of only a pilot's license is evidence, or the basis of a presumption, that the possessor is lacking in skill as a [commercial] pilot. This does not *necessarily* follow. He *may* not have chosen to take further examinations although he *might* have been abundantly qualified for commercial or transport license for aught the record shows. . . ." 138 F.2d at 341 (emphasis added). Why does the court insist on talking about *possibilities* which are patently less than probable? Smith v. Whitley, 223 N.C. 534, 27 S.E.2d 442 (1943), is similar. The classic opinion is that in Brown v. Shyne, 242 N.Y. 176, 151 N.E. 197 (1926).

---

[42] Cf. Martin v. Herzog, 228 N.Y. 164, 126 N.E. 814 (1920) (fact that unlighted vehicle was struck by another in the dark is evidence that the collision occurred *because* of the lack of lights). Dean Green agrees with the position stated in the text. See discussion of Mundy v. Pirie-Slaughter Motor Co., 146 Tex. 314, 206 S.W.2d 587 (1947), in Green, Proximate Cause in Texas Negligence Law, 28 Texas L. Rev. 755, 766 (1950), and see Green, Proximate Cause in Connecticut Negligence Law, 24 Conn. B.J. 24, 32 (1950). If the rule suggested here should emerge, it might well be qualified by holding that the inference of negligence from want of a license is rebutted by a showing that the actor was qualified for a license, so far as skill was concerned.

The lack of registration for an automobile involved in an accident evokes varied responses from different courts. In the great majority of jurisdictions the driver of an unregistered automobile is not liable to those he injures if he is otherwise exercising due care. Gilman v. Central V. Ry. Co., 93 Vt. 340, 107 Atl. 122 (1919); Armstrong v. Sellers, 182 Ala. 532, 62 So. 28 (1913). See annotation, 163 A.L.R. 1375 (1946) (citing earlier annotations), and cases cited by Carpenter, Workable Rules for Determining Proximate Cause, 20 Calif. L. Rev. 396, 410 n.133 (1932). This result should be reached not, as many courts reason, because the violation of the statute is not the proximate cause of the injury (see, e.g., Gonchar v. Kelson, 114 Conn. 262, 158 Atl. 545 (1932)), but because the plaintiff is not within the scope of duty owed by the defendant with respect to registration. See discussion of the Gonchar case by Green, Proximate Cause in Connecticut Negligence Law, 24 Conn. B.J. 24, 28–30 (1950). See further discussion of this point in the text, §20.5 at note 18 *infra.*

Massachusetts is probably the only state which still considers the unregistered vehicle an outlaw or trespasser on the highway to which no duty is owed, Chase v. N.Y. Cent. & H.R.R. Co., 208 Mass. 137, 94 N.E. 377 (1911), and which is responsible for all harm, regardless of negligence. McDonald v. Dundon, 242 Mass. 229, 136 N.E. 264 (1922). For some of the ridiculous bases upon which an automobile has been found to be unregistered in Massachusetts, see Comment, 10 B.U.L. Rev. 211 (1930). But Massachusetts has apparently come to the point where it is at least unwilling to extend this rule to new situations. See Galbraith v. Levin, 323 Mass. 255, 81 N.E.2d 560, 563 (1948), in which the owner was held *not* liable for injuries inflicted by a thief who was negligently driving his car which was unregistered and in which, in violation of a statute, he had left the keys. This decision directly overruled, Malloy v. Newman, 310 Mass. 269, 37 N.E.2d 1001 (1941), which involved very similar facts and which rested heavily on the fact that the automobile was unregistered.

CORRELATION AND CAUSATION

The discussion by Mueller and Schuessler concerning correlation notes that correlation does not necessarily indicate causation. At the same time, it is true that all observers of human behavior, including lawyers and social scientists, engage in the analysis of causal relationships and depend upon correlation to do so. And correlations can lead to inferences about the operation of causation--although they cannot prove it in a final sense. The task, then, is to place correlation in its proper perspective with regard to causation; and we feel that the following reading provides an excellent discussion of the problem.

From An Introduction to Logic and Scientific Method by Morris R. Cohen and Ernest Nagel, copyright, 1934, by Harcourt, Brace & World, Inc.; C 1962, by Ernest Nagel and Lenora Cohen Rosenfield. Reprinted by permission of the publishers.

Morris R. Cohen and Ernest Nagel, An Introduction to Logic and the Scientific Method (New York: Harcourt, Bruce & World, Inc., 1934), pp. 245-272.

§ 1. TYPES OF INVARIANT RELATIONS

The search for order among facts is a difficult task. Few succeed in it. But it has always been the hope of some thinkers that easily learned rules might be found according to which anyone undertaking such a task may be assured that success will crown his efforts. And some writers on scientific method have proudly believed they had actually found such rules. Francis Bacon was one of them. "Our method of discovering the Sciences," he wrote, "is such as to leave little to the sharpness and strength of men's wits, but to bring all wits and intellects nearly to a level. For as in drawing a straight line, or describing an accurate circle by the unassisted hand, much depends on its steadiness and practice, but if a rule or a pair of compasses be applied, little or nothing depends upon them, so exactly is it with our method."[1] The methods which Bacon recommended for discovering the causes of things are popularly believed to express the nature of scientific method. They were elaborated by John Stuart Mill and formulated by him as the *methods of experimental inquiry*.

Before examining these methods, let us note some preliminary difficulties. In Chapter XI we tried to indicate that isolated facts do not constitute a science, and that the object of science is to find the order among facts. But what kind of order? It is commonly believed that science is interested exclusively in *causal* order. The analysis of the meaning of "causality" is a most difficult task. We cannot enter upon it here, for it is not a task for the logician. But we must observe that various kinds of order are sometimes confounded as identical with the causal order. We may then be prepared to state the general character of the *sort* of order for which the sciences are in search.

---

[1] *Novum Organum*, Bk. I, LXI.

1. There is a type of order so familiar that it generally escapes notice. All of us recognize certain things as water, other things as wood, still others as steel, and so on. Why do we assign special names such as "water," "wood," and "steel"?

We apply the name "water" to something which is a liquid when above a definite temperature, a vapor when above another temperature. This "substance" is generally translucent, odorless, colorless; it has a constant density, and is practically incompressible; it quenches fire and thirst. "Water" denotes a *constant conjunction of properties,* and a name is given to it in order to distinguish it from other such conjunctions or "things." So also for "wood" and "steel."

The vague concept of "thing" denotes, therefore, a very elementary but fundamental type of order. It denotes a certain invariable conjunction or association of properties that is different from other conjunctions. Such a type of order would probably never be called "causal." But the discovery of this kind of order is fundamental for the discovery of any other kind of physical order. Different kinds of things have been recognized in the most primitive stages of man's history. But the process of classifying and cataloguing our experiences is not complete, and perhaps never will be.

2. A type of order frequently recognized is one involving a temporal span or direction. "Iron rusts in moist air," expresses one instance of such an order. It is this kind of order that is generally regarded as causal.

The "common-sense" notion of cause is an interpretation of nonhuman behavior in terms peculiarly adopted to human behavior. Thus, "John broke the window," is supposed to express a causal relation, because there is an agent "John" who *produced* the breaking of the window. So in the proposition above, the moist air is said to be the cause, of which the rusting iron is the effect. And the moist air is said to be the cause because it is believed to *produce* the rusting. In the popular mind, all *changes* require *causes* to explain them, and when found are interpreted as agents producing the change.

It is very difficult to make clear what is understood by causes "producing" their effects. When we reflect on the matter, all that we seem capable of discovering in alleged instances of causality, is an instance of an *invariable relation* between two or more *processes.* It is not the mere *existence* of John which is the cause of the broken window. What is important is the constant relation which holds between a certain kind of behavior of John and a certain other kind of behavior of the glass. One of the characteristics of the causal relation as generally understood is that it is asymmetrical and temporal in nature.

But even "common sense" soon recognizes that the apparent invariability of alleged causal relations is often specious. Iron does not *always* rust in moist air, and a window is not *always* broken when a brick is hurled at it. Even "common sense" discovers that other factors must be present in these situations besides those already noted. Hence it is not moist air alone that is the cause of rusting. Those other factors are then sought for which seem to be necessary for the occurrence of the effects. In this way there is a gradual transition from the crude and approximate uniformities observed in everyday experience, to the more completely analyzed invariant relations of the developed sciences.

3. Many uniformities are expressible by numerical equations. Ohm's law in electricity states that the current is equal to the potential difference divided by the resistance. The principle of the lever states that equilibrium is obtained when the two weights vary inversely to their distances from the fulcrum.

Invariable relations of this type no longer assert a sequence in time, and they are probably never regarded as illustrations of causal order. It is true that in making experiments upon an electric circuit we may alter the current first and subsequently note the change in the potential difference. But what Ohm's law states is not the order in which we make observations; it states that the measurable elements observed stand to each other in the specified invariable relations.

4. A fourth type of order is illustrated in such comprehensive theories as the theory of gravitation, or the kinetic theory of matter. In such theories, not all the elements between which the invariant relations are asserted to hold are directly observable. Nor are all the relations which are asserted to hold between the elements capable of direct experimental confirmation. Thus the atoms, their motions and collisions, the invariance of their average energies, cannot be verified directly.

The function of such comprehensive theories, as we have already seen, is that they enable us to show that many numerical as well as qualitative laws which are in fact experimentally confirmable, are not isolated from one another. Such laws can often be shown to be the necessary consequences of the more abstract and inclusive order asserted in the theory. Thus the numerical relations between the temperature, the volume, and the pressure of gases; the numerical laws connecting the density and the specific heat of gases; the relations between melting-point, pressure, and volume of solids—these are all derivable by logical methods from the assumptions of the kinetic theory of matter.

If we examine these four general types of order, we discover that a generic feature of all of them is the assertion of some kind of invariable relation between various kinds of elements. The relation in some cases may involve a temporal asymmetry; the cause is popularly said to precede the effect in time. In other cases the temporal reference is missing. It is the *invariability* which seems to be significant, both theoretically and practically.

By the *cause* of some *effect* we shall understand, therefore, some appropriate factor invariably related to the effect. If A *has diphtheria at time t* is an effect, we shall understand by its cause a certain change *C*, such that the following holds. If *C* takes place, then *A* will have diphtheria at time *t*, and if *C* does not take place, *A* will not have diphtheria at time *t*; and this is true for all values of *A*, *C*, and *t*, where *A* is an individual of a certain type, *C* an event of a certain type, and *t* the time.

The search for "causes" may therefore be understood as a search for some invariable order between various sorts of elements or factors. The *specific* nature of this order will vary with the nature of the subject matter and the purpose of the inquiry. Moreover, the *specific* nature of the elements between which the order is sought will also differ for different inquiries. In some cases we already know the invariable order and some of the elements, and then our search is for one or more further elements. Thus, finding a person dead from wounds and knowing the conditions under which such a death results, we look for the murderer. In other cases, we may know the elements and search for an invariable order between them. Thus we may note hot water being poured into a glass and also note that the glass cracks; we may then look for the structure of the relations connecting two such processes. In still other cases, we may notice some change and then look for other as yet unknown changes connected with it in some ways not yet known. Thus we may note the aurora borealis, and then search for the circumstances with which it is connected in some fashion or other.

The kind of elements or changes for which we look depends on the structure of the order in which we are interested. The answer to the question, "Who killed the Archduke Ferdinand at Sarajevo?"

must be of the form: "The person or persons *A, B, C* and so on, are the assassins of the Archduke." On the other hand, the question, "What killed the Archduke?" must be answered according to the kind of specific order for which we are in search, and according to the purpose of the inquiry. One answer may be, "A certain revolver was the cause of his death." Other possible answers are: "Certain social and political conditions were the cause of his death"; "The cessation of the oxygen supply to the cells of his body was the cause of his death." In other words, the kind of order, as well as the kinds of elements we look for, is determined by the nature of the problem which generates the inquiry. What is an adequate answer to one question will not, in general, be adequate to another.

In the light of the great variety in the kinds of specific orders and factors which may be the objects of an inquiry, it may seem preposterous to believe that any general rules can be stated which will enable us to find satisfactory answers to all possible problems. We shall not prejudge the matter, however, and shall examine at length the experimental methods formulated by Mill.

§ 2. THE EXPERIMENTAL METHODS IN GENERAL

The "experimental methods," according to their author, have a twofold function to fulfill. They are, in the first place, *methods of discovering* causal connections. Mill believed that by employing the methods the order in which facts stand to one another can be found. Against his critics he maintained that all inferences from experience are made by means of these methods. If "no discoveries were ever made by the . . . methods . . . none were ever made by observation and experiment; for assuredly if any were, it was by processes reducible to one or other of those methods."[2] It is these methods which supply the first generalizations upon which all subsequent construction of hypotheses depends.[3]

But secondly, the methods have a *demonstrative* function as well. Mill conceived the business of logic to be concerned with proof. Inductive logic, according to him, should supply "rules and models (such as the Syllogism and its rules are for ratiocination) to which if inductive arguments conform, those arguments are conclusive, and not otherwise."[4] The methods are therefore to be *tests* of any

[2] *A System of Logic*, Vol. I, p. 500.
[3] *Ibid.*, p. 501.
[4] *Ibid.*, p. 500.

experimental procedure. Just as the evidence for a proposition is conclusive if the relations between the propositions offered in evidence and the one in question conform to the conditions for necessary inference, so an "inductive argument" is valid, it is claimed, if it conforms to the "experimental methods." The conclusion of investigations into matters of fact could therefore be *absolutely certain.*

We shall be able to evaluate these two claims more clearly if we state what the general nature of the methods is, and if we recall what are the conditions under which an inquiry can be carried on. In the first place, some selected portion of our experience is taken for further study because of its problematic character. The problem must then be formulated in terms of the situation which provokes the inquiry, and an analysis of the situation must be made into a certain number of factors, present or absent, which are believed to be relevant to the solution of the problem. Now the order for which we are in search is expressible, as we have seen, in the form: C is invariably connected with E. And this means that no factor can be regarded as a cause if it is present while the effect is absent, or if it is absent while the effect is present, or if it varies in some manner while the effect does not vary in some corresponding manner. The function of experiment is to determine with regard to each of the factors entertained as a possible cause, whether it is invariably related to the effect. If C and E are two factors or processes, there are four possible conjunctions: we may find either CE, $C\bar{E}$, $\bar{C}E$, or $\bar{C}\bar{E}$, where $\bar{C}$ and $\bar{E}$ denote the absence of these factors. To show that C is invariably connected with E we must try to show that the second and third alternatives do not occur.

When the problematic situation is complex, and contains distinguishable factors as components, we can establish an invariable relation between the effect and some of the possibly causal factors only by showing that such factors do or do not meet this *formal* condition for invariable connection. It is necessary, therefore, to vary the supposedly relevant factors *one at a time,* and consequently to analyze the situation into factors that are relatively independent of one another.

The function of experiment, as we shall see, is *eliminative.* And the methods of experimental inquiry, as we shall also see, have precisely that function.

§ 3. THE METHOD OF AGREEMENT

The Method of Agreement as a Canon of Discovery

Whether we are searching for the cause of some event or for the effect, we begin with a situation which can be identified as *one of a type*. Suppose that one morning we find the flowers in the gardens of a suburban town to have withered over night. How shall we proceed to find the cause?

The first experimental canon instructs us as follows: *If two or more instances of the phenomenon under investigation have only one circumstance in common, the circumstance in which alone all the instances agree, is the cause (or effect) of the given phenomenon.*[5] The phenomenon we are investigating is the withering of flowers; the withered flowers in the several gardens are the instances. We must, accordingly, examine the instances for the common circumstances or factors. We note that the gardens differ in many ways: in the quality of the soil, in the kinds of flowers they grow, in their size, in their location, in the character of their gardeners. We note also that the temperature had fallen very sharply during the night. We conclude on the basis of the canon that the drop in temperature is the cause of the withering of the flowers. Why are we justified in drawing this conclusion? Why do we not say that the quality of the soil is the cause? Because, it may be said, the factors which are absent when the phenomenon is present cannot be invariably connected with the phenomenon: the quality of the soil is *not* the same in each instance of the withered flowers. Hence the invariable circumstance of the drop in temperature must be the cause. And indeed sharp changes in temperature are known to be fatal to flowering plants. The canon seems to be a successful method, therefore, both for discovering and for "proving" causes.

Unfortunately, in this illustration we knew what the cause of the withering of the flowers was *prior* to the application of the method. It is not surprising that we have been able to discover the cause. We must employ the canon to discover the cause of some phenomenon for which we do not happen to know the cause antecedently.

Now baldness in men is a phenomenon for which the cause is not known. If the canon is an effective instrument of discovery, no student of logic need suffer long from a naked scalp. In accordance with the canon, we find two or more bald men and search for a single common circumstance. But immediately we run into great difficulties. The method requires that the men should differ in all respects except one. We will have rare luck indeed (or we are perhaps rather unimaginative in noting common circumstances) if we succeed in obtaining a collection of men who satisfy this condition. (If we are not too particular about the nature of the "common" circumstance, we may perhaps hold that the common circumstance to all bald men is their being organic bodies.)

---

[5] *A System of Logic*, Vol. I, p. 451.

Let us waive this point. We then meet a more serious obstacle. How shall we go about identifying the common circumstance or circumstances? If one specimen has blue eyes, must we examine all the others for color of eyes? If one of the bald men confesses to having been brought up on cod liver oil, must we discover whether the others were brought up the same way? But the number of such circumstances to which we might be directed is without limit. Date of birth, books read, food eaten, character of ancestry, character of friends, nature of employment, are some examples of circumstances to which we might pay attention. If, therefore, the *common* circumstance can be found only by examining all the instances for *every* circumstance which some one or other of the bald group may possess, we can never find *all* the common factors in this way. We can carry on the search for a common factor only if we disregard most of the circumstances which we may find as not *relevant* to the phenomenon of baldness. We must, in other words, start the investigation with some hypothesis about the possible cause of baldness. The hypothesis which selects some circumstances as possibly relevant and others as not is constructed on the basis of previous knowledge of similar subject matter. This hypothesis is not supplied by the canon. Without some hypothesis on the nature of relevant factors the canon is helpless to guide us to our goal.

We have been pretending that the circumstances or factors which are present in an instance are distinct from one another, and that each comes labeled, as it were, with a tag saying, "I am a circumstance." But it simply is not true that an instance of a phenomenon is presented to us as a unique set of sharply defined factors, immediately recognizable by us as such, so that each factor can be examined and varied independently of any other. Now the method of agreement requires a comparison of circumstances in two or more instances. Unless, therefore, an analysis of an instance into its factors could be made *prior* to the use of the method, the method would be altogether useless.

How do we divide an instance into its factors, and is every analysis of instances into factors equally valid? Consider the following experiment. In two or more test tubes of unequal sizes and each filled with a liquid of a different color, a precipitate is formed. We wish to determine the cause of the precipitation and we find that each instance can be analyzed into the following circumstances: (1) A test tube of a certain size (2) contains a liquid with a specific color (3) to which sulphuric acid has been added; (4) precipitate is formed. Factors 1 and 2 are eliminated by the canon, since the instances vary with respect to these; the canon fixes upon the addition of the acid as the cause of the phenomenon (4). But this is not the *only* way in which the instances could have been analyzed. We may have tried the following division into circumstances: (1′) A test tube of a certain size (2′) contains a liquid (3′); sulphuric acid is added to different-colored liquids; (4′) precipitate is formed. On the basis of this analysis, the method would conclude that the factor (2′) is the cause of the precipitation. This conclusion is in fact false. It is false because the second method of analyzing the instances is not a "proper" one.

Let us consider another illustration. We wish to know the cause of headaches. We find that it was preceded in some cases by eyestrain, in some cases by indigestion, and in some cases by the hardening or other disturbances of certain blood vessels. If we take this canon literally, none of these are common circumstances and therefore to be considered as causes. This, however, would be an error, due to an inadequate analysis of what is a headache and of what factors in all these mentioned circumstances are relevant to the different kinds of headaches. A greater refinement of the causes must be accompanied by an equal refinement of the effects. If we ask what is the cause of disease, we have grouped a large number of phenomena under one rubric and the diverse causes of the different kinds of disease must be similarly grouped. This point will be discussed more fully under the heading of The Plurality of Causes.

It follows that not every analysis into factors is equally valid. It is not valid because *in the light of our knowledge* we must not separate, in such experiments as the one above, the volume and place occupied by a liquid from what has been added to it. Now the method of agreement cannot inform us which is the proper analysis. It cannot discover for us how to divide instances into factors such that invariable relations can be found to hold between some of them. The method cannot possibly function unless, once more, assumptions about relevant factors are made.

Morris R. Cohen and Ernest Nagel

The Method of Agreement as a Canon of Proof

Let us consider next whether the method of agreement is one of proof, even if we must recognize that it is not one of discovery. Does it follow that because a search for the cause of a phenomenon conforms to the conditions stated by the method, the conclusion of the search is thereby *demonstrated* to be true? That it does not so follow can be easily seen. The cause of a phenomenon must be invariably related with the phenomenon. But we cannot examine more than a *limited* number of instances of any alleged invariable relation. Even if we could be absolutely certain that the circumstance claimed as a cause is the *single* common circumstance, can we be certain that it is *invariably* (for an *unlimited* number of instances) connected with the phenomenon? Thus we may find in a very large number of instances of typhoid that the activity of microörganisms is the one common factor that is present. It does not follow that this factor is *always* present in the still unobserved instances of typhoid. Not every actual conjunction of circumstances is *indefinitely* repeated.

The reader may perhaps believe that the inference from the observed conjunction of factors to an invariable conjunction is legitimate in virtue of the "uniformity of nature." We shall not disturb the reader's faith in this familiar doctrine at this point. But, as we shall see presently, such a faith has no evidential value in demonstrating the existence of invariable connections.

Not only will the method not serve to prove the presence of a causal relation; it may, on the contrary, lead us to affirm some factor to be the cause when it is not. We have seen this already in connection with the problem of analyzing instances into circumstances. We can see this in another way. Suppose a professor of hygiene finds that he had a splitting headache on three successive nights. He recollects that on Monday he read for ten hours and then took a walk; on Tuesday, he found the dinner delicacies irresistible, ate too much, and then sought repentance by taking a walk; on Wednesday, he slept during the day and then sought refreshment in a walk. If he were to employ the method of agreement, he might conclude that walking was the cause of his headache. But this is quite contrary to fact, since the walks he took (we happen to know on other grounds) have nothing to do with bringing on the headaches. A false conclusion was drawn by using the method, since the instances to which it was applied were not analyzed properly into the right circumstances.

This illustration suggests another familiar doctrine. The method of agreement does not provide a "water-tight" proof, it is claimed, because there is such a thing as a plurality of causes. The same phenomenon is not always produced by the same cause, Mill believed: "There are often several independent modes in which the same phenomenon could have originated."[6] A house may be destroyed by fire, or by an earthquake, or by cannon fire. Consequently, this method cannot find *the* cause. It was such a reflection which compelled Mill to recognize an imperfection in the canon of agreement and which led him to supplement it with the canon of difference. We shall return to the doctrine of plurality of causes presently. Whether the doctrine is tenable or not, its formulation and adoption by Mill show the need for some criterion of the correct analysis of instances into circumstances. It is a criterion not supplied by the canon.

Employing the canon does not guarantee that *all* the necessary conditions for the occurrence of the phenomenon will be found. Why is the mercury column in barometers generally around thirty inches high? If we employ this method we may conclude that since there is a vacuum at the top of each column, the existence of the vacuum is the cause of the observed height of the mercury. This is a mistake, since we know that the occurrence of a vacuum is not a sufficient condition for the height of the column. The atmospheric pressure, the temperature of the room, are other conditions which are indispensable in order to explain the height of the mercury. The method of agreement may, therefore, overlook certain general conditions which must obtain. It may fix our attention only upon certain obvious, even if necessary, features of the instances.

The Value of the Method of Agreement

The method of agreement is therefore useless as a method of discovery and fallacious as a canon of proof. Has it then no value? It has a limited value, if stated negatively: *Nothing can be the cause of a phenomenon which is not a common circumstance in all the instances of the phenomenon.* Thus stated, it is clearly a method of eliminating proposed causes which do not meet the essential requirements of a cause. A circumstance that is not common to all instances of a phenomenon cannot, by definition, be causally related to it.

A search for causes begins with some assumptions about the possibly relevant factors. Thus in studying baldness we may begin as follows: Baldness is due to congenital, hereditary factors, or to the characters of the diet, or to the nature of the headwear, or to some previous disease. The method of agreement helps to *eliminate* some or all of the suggested alternatives. We may discover that the character of the food eaten by bald men is not a common feature; and according to the principle of *tollendo ponens* we can conclude therefore that only the three other alternatives remain to be examined, that is, baldness is either congenital, or it is due to the nature of the hats worn, or it is due to some previous disease. We may proceed in this fashion until we have eliminated all the suggested alternatives, or found one or more which cannot be eliminated.

---

[6] *A System of Logic*, Vol. I, p. 505.

Unless, however, we have been fortunate enough to include the circumstance which is in fact the cause of the phenomenon in the enumeration of alternatives, the method of agreement can never identify the cause. Its function is to help *eliminate irrelevant circumstances.*

§ 4. THE METHOD OF DIFFERENCE

The Method of Difference as a Canon of Discovery

The method of agreement was recognized as faulty by Mill because we cannot be certain that the phenomenon investigated has only one cause. It was believed to be useful in those cases where we could not *alter* the circumstances at will. Hence, it was regarded primarily a method of *observation* rather than *experiment.* It was believed, however, that the shortcomings of this method can be overcome by the use of a second canon, the *method of difference.*

This second method requires two instances which resemble each other in every other respect, but differ in the presence or absence of the phenomenon investigated. Its full statement is: *"If an instance in which the phenomenon under investigation occurs, and an instance in which it does not occur, have every circumstance in common save one, that one occurring in the former; the circumstances in which alone the two instances differ, is the effect, or the cause, or an indispensable part of the cause, of the phenomenon."*[7]

Let us see whether this canon is effective in discovering causes. Suppose the reader buys two fountain pens of like make, fills them with the same kind of ink, places them in his pocket, and takes a long walk before he sits down to write. When he does, he discovers that one of the pens leaks. What is the cause of this? It seems as if the conditions for applying the canon are all present. The pens are alike, but one leaks and the other does not. And if the reader employs the canon he may "discover" that the rubber sack in one of the pens has lost its elasticity and is extremely porous; the other pen has not this defect. The condition of the rubber, the reader may conclude, is the cause of the leak.

---

[7] *A System of Logic,* Vol. I, p. 452.

But is the matter as simple as that? If we take the canon seriously, we must conclude that the method cannot be applied in this inquiry, since it requires that the two pens be *exactly* alike in all circumstances except those mentioned. But the two pens do differ in very many ways: one was made before the other, or by a different workman; the shapes of the pens have minute differences; chemical analysis reveals other differences; the pens were not placed in exactly the same position in the reader's pocket, nor were they warmed equally by his body.

If anyone should object that the pens need not be *exactly* alike, but only alike in *relevant* factors, we must reply that it is precisely such judgments of relevance which are required before the canon may be used; and that the canon does not supply this vital information. If, however, the objector should declare that the two pens can be shown to be alike by examining *all* the circumstances, we would be forced to reply that an exhaustive examination of the circumstances is impossible; and that if it were possible, the canon would be unnecessary to discover the factor which is present when the phenomenon is present and absent when it is absent.

This canon, like the previous one, requires, therefore, the antecedent formulation of a hypothesis concerning the possible relevant factors. The canon cannot tell us what factors should be selected for study from the innumerable circumstances present. And the canon requires that the circumstances shall have been properly analyzed and separated. We must conclude that it is not a method of discovery.

The Method of Difference as a Canon of Proof

Is it a method of proof or demonstration? No more than the method of agreement! Whatever value the canon may seem to have depends on the assumption that differences are noted with respect to the presence or absence of a *single* factor. But can the canon assure us that the factor is not *complex*?

Suppose a man is psychically and socially maladjusted. He suffers from uncomfortable dreams. He goes to a psychoanalyst, who persuades him to disclose intimate details of his autobiography, and in particular of his sex life. The man "recovers" from the maladjustment and the dreams cease. The canon seems to be applicable: the "single difference" in the events of the man's life during this period is the expression of his hidden sex desires. Can we validly conclude that the man's talking freely on sex is the cause of his recovery? Certainly not. The change in the man's life may in fact be due simply to his finding in the analyst a *sympathetic audience* on any subject whatever, or it may be due to the cessation of certain organic disturbances unknown to the patient or the analyst.

Can this method demonstrate an *invariable* connection by an examination of *two* instances? Suppose the reader spends a sleepless night, but on the following night he rests peacefully. The reader may be able to convince himself that the single "significant" difference in his behavior on the two days preceding these nights is that he had drunk coffee on the first day but not on the second. Can the reader validly conclude that drinking coffee is (for him) the invariable cause of sleeplessness, other things being equal? It may be true that his sleeplessness on the first night was in fact due to the coffee. Nevertheless, it may be that it is not drinking *coffee as such* which produces the undesirable result. The insomnia may be due to a *drug* that the coffee contained. The reader's sleeplessness on that particular night was, by hypothesis, due to drinking that particular coffee. But it does not follow that in general coffee-drinking is followed by a restless night. The application of the canon does not, therefore, uniformly lead to the detection of the factors in terms of which an invariable relation can be expressed, and it may lead to an affirmation of an invariant relation where none in fact exists. The canon does not safeguard us against the fallacy known as *post hoc, ergo propter hoc:* sleeplessness may *follow* drinking coffee, but sleeplessness may not occur *because* coffee was drunk.

The statement of the canon clearly recognizes that the factor noted by it may be only a *part* of the cause. This is a very important qualification. The invariable relations for which the sciences seek are such that if a determinate set of circumstances are present, some other circumstance will always accompany these. The discovery of a *partial* set of circumstances is often not enough. Now the method of difference cannot guarantee that the sufficient con-

POOR ORIGINAL COPY - BEST
AVAILABLE AT TIME FILMED

ditions for a phenomenon have been found. We cannot infer that rain is a *sufficient* condition for the rich harvest in one part of the state on the ground that there was a drought in another part, even though the quality of the seeds planted and of the soil and the quantity of sunshine were in all significant ways the same. For it isn't the *rain alone,* but the rain *together with* the soil, seeds, and sunshine, which provides the adequate conditions for a bumper crop. The canon may therefore very easily direct our attention to extremely partial and even superficial factors in the complete situation. On the basis of the canon we might argue that since conditions in Europe in January, 1914, were the same as in July, and since the only relevant difference was the murder of the Archduke Ferdinand, the assassination was the cause of the World War. Without denying the importance of this event in explaining just when the war did take place, no serious student of affairs would hesitate to point out the complicated national, diplomatic, and socioeconomic factors which were part of the conditions required to explain the occurrence of the war.

The Value of the Method of Difference

The method of difference cannot, therefore, be regarded either as a method of discovery or as a method of proof. But, like the method of agreement, it has a limited value when stated negatively: *Nothing can be the cause of a phenomenon if the phenomenon does not take place when the supposed cause does.* Thus stated, it is clearly a method of eliminating one or more proposed causes which do not meet the essential requirement of a cause. A circumstance that is present whether the phenomenon is present or not cannot, by definition, be causally related to it. Thus if we are studying rheumatism, we may entertain the hypothesis that it is caused by excessive starch in the diet, or by lack of exercise, or by a focal infection in the teeth. Provided that these alternatives represent an adequate analysis and separation of the circumstances, we may be able to eliminate the diet theory of rheumatism if we can show that large quantities of starch can be consumed without being followed by the effect. Proceeding as we did with the canon of agreement, we may, in accordance with the principle of *tollendo ponens,* be able to eliminate all but one alternative. Again, the method of difference is helpless if we have not the sagacity to include in the alternatives considered for further study the factor which is in fact the cause.

§ 5. THE JOINT METHOD OF AGREEMENT AND DIFFERENCE

The two methods so far considered require conditions for their application which we can never find realized. The first method requires instances which are unlike in every respect except one; the second requires instances which are alike in every respect except one. When the phenomenon is dependent upon a complex set of conditions, it is difficult to separate the factors involved and vary them one at a time. Mill therefore proposed a combination of the two preceding methods. Its formulation is: "*If two or more instances in which the phenomenon occurs have only one circumstance in common, while two or more instances in which it does not occur have nothing in common save the absence of that circumstance; the circumstance in which alone the two sets of instances differ, is the effect, or the cause, or an indispensable part of the cause, of the phenomenon.*"[8]

The statement of the canon, however, is really absurd. According to it we require two *sets* of instances. In one set the phenomenon occurs, and the instances *taken together* must have a single common circumstance, although taken two at a time they may agree in more than one circumstance. In the second set the phenomenon does not occur, and the instances must be so chosen that they have *nothing* in common, taken together, save the absence of the phenomenon. But if we follow these instructions we may include *anything we wish* in the second set, since the *absence* of some one character need be the only identical feature in all of them! Suppose we wish to discover the conditions which make for divorce. According to the method, we must in the first place examine a number of divorced couples, and in the second place, examine a number of cases where divorce has not occurred, for example, among flowers, children, mountains, bachelors, and so on. We could not possibly use *these* negative instances to determine the cause of divorce. We must therefore modify the formulation of the canon. The negative instances must be all of a type in which the phenomenon is *capable* of being present when the adequate conditions are supplied.

As a method of discovery and proof, this canon combines all the defects of the first two canons; its virtues are the virtues of either. However, it does formulate certain aspects of methods employed in making comparisons between large groups. If we were to try the method of difference alone for finding the cause of divorce, we would require two couples, one divorced, the other not, that are alike in *every* way except one. This is hardly feasible. If, however, we were to examine a large number of married pairs, we might be able to show that some of the circumstances which are common to all of them are not *significant* for their continuing in the married state, provided we could also show that divorced pairs show the *same* common features. We might be unable to identify the cause of divorce by this method. Nevertheless, by examining several large groups we might be able to show some relation between the relative frequency of divorce and such factors as differences in the age, education, health, and so on of the parties to a marriage. Such statistical information may be all that can be obtained. The knowledge of the relative frequency of divorce for individuals differing considerably in age, for example, will be useless for determining whether divorce will terminate the marriage of some particular married pair. It may be very useful in ascertaining how often we may expect divorces in a very large group.

---

[8] *A System of Logic,* Vol. I, p. 458.

§ 6. THE METHOD OF CONCOMITANT VARIATION

The elimination of irrelevant circumstances, which we have seen to be the function of the preceding canons, cannot be performed by them in all cases. For it is sometimes impossible to exclude or isolate the cause completely. If we wish to find the cause of the rise and fall of the tides of rivers and seas, we cannot use the canon of difference, since we cannot find an instance in which such a body of water does not show the phenomenon of tidal behavior. We cannot show with the method of difference that the sun and moon are the cause of tides, since we cannot eliminate the action of these bodies in any instances. And we cannot use the canon of agreement, because we cannot remove from the instances of tidal behavior such ineradicable common circumstances as the presence of the fixed stars.

In such cases, however, we may notice or introduce variations in the degree or magnitude of the effect, and find a corresponding variation in some circumstance, without thereby *completely* eliminating either the effect or the supposed cause. The method of concomitant variation has been formulated by Mill to deal with such phenomena. Its statement is: "*Whatever phenomenon varies in any manner whenever another phenomenon varies in some particular* *manner, is either a cause or an effect of that phenomenon, or is connected with it through some fact of causation."*[9]

This canon can be employed, therefore, only if degrees or magnitudes of effects and causes can be distinguished. The previous canons are *qualitative* methods, since their use requires simply the determination of the presence or absence of some character or quality. The present canon is *quantitative,* and requires the aid of measurement and statistical technique.

The Canon of Concomitant Variation as a Method of Discovery

An examination of the statement of the canon of concomitant variation must make us suspicious of its efficacy as a method of discovery. It declares that if a phenomenon varies in any manner *whenever* another phenomenon varies in some manner, a causal relation is present. Now if the concomitant variation is actually invariable, and this seems to be required by the word "whenever," a causal relation is indeed present. But if in order to employ the canon we must know antecedently that the mode of variation is invariable, of what use is the canon? We do not, in that case, need the canon to discover for us the cause. And the canon by its own admission is perfectly helpless in finding the rule of variation or in demonstrating that a supposed mode of variation is invariable.

This suspicion is strengthened if we try to use the canon. Suppose we notice that the temperature in a region varies in some determinate manner during several months. What is the cause of this variation? We look for some circumstance present during these months which also undergoes some variation. But which circumstances shall we examine? Certainly not *all* the circumstances, not even all the varying circumstances. The formulation of hypotheses and judgments of relevance are required before this canon can be employed.

The complicated causal dependencies between several variables which the natural sciences study cannot possibly be unraveled unless hypotheses based on knowledge of mathematical relationships are formulated concerning them. Even such a relatively simple rule of variation as the inverse-square law of gravitational attraction cannot be obtained by merely observing the behavior of planets.

The Method of Concomitant Variation as a Canon of Proof

The mere presence of a concomitant variation of temperature and some other factor is not sufficient to establish a causal connection. Suppose the changes in the daily temperature in New York City during one year could be shown to vary with the daily death rate in China in that period. Such correlations would be regarded by most competent judges as fortuitous, because they have some prior knowledge concerning the relevant factors in the production of temperature changes. Even very high correlations, especially in the social sciences, do not necessarily signify an invariable connection. For the phenomena between which such correlations can be established may be in fact unrelated in any way which would warrant our believing them to be invariably connected. A little statistical skill and patience make it possible to find any number of high correlations between otherwise unrelated factors. We do not discover causal connections by first surveying all possible correlations between different variables. On the contrary, we suspect an invariable connection, and then use correlations as corroborative evidence.

---

[9] *A System of Logic,* Vol. I, p. 464.

Moreover, the correlations obtained on the basis of an examination of a finite number of pairs of variables are unreliable, because we cannot be sure that the rule of variation remains the same outside of the actually observed limits of variation. We may, by good fortune, come to the study of gases by selecting a gas like helium at a high temperature. We may then observe that if the temperature remains constant, the pressure varies inversely as the volume of the gas. We may observe this rule of variation for certain intervals of temperature, and then extrapolate the rule for every value of the temperature, or even for any gas at any constant temperature. But if we do so, we are sure to blunder, since it is now known that Boyle's law is true only for a few gases under ideal conditions. Indeed, a rule of variation which has been found to hold within certain intervals may become altogether inaccurate outside those limits, not only because the rule of variation is different, but also because circumstances negligible within those limits cannot be neglected outside that interval. The period of a pendulum is proportional to the square root of its length if the arc of the swing is small. When the arc of the swing is increased, the period (theoretically and approximately) is still related in this way to the length; nevertheless, the factor of air resistance must now be considered, so that the period can no longer be rendered by that simple formula.

The Value of the Method of Concomitant Variation

The method of concomitant variation cannot therefore be accepted as a method of either discovery or proof. Its value lies partly in suggesting lines of inquiry for causal relations and in helping to corroborate hypotheses of causal connection. Its chief value, however, is to help eliminate irrelevant circumstances. For nothing will be regarded as the cause of a phenomenon if when the phenomenon varies that thing does not, or when the phenomenon does not, that thing does. Consequently, the method will help eliminate those factors suggested by the hypothesis guiding the inquiry which do not conform to this condition. Mill's statement of the canon asserts that if *C* varies whenever *E* varies, *C* and *E* are causally related. We have seen that this claims too much. All that can be affirmed is that *C* and *E* are *not* causally related if *C* and *E* do not vary concomitantly. And even in this modified form the method will not save us from error if the circumstances denoted by *C* and *E* are not properly analyzed.

§ 7. THE METHOD OF RESIDUES

The remaining method of "discovery and proof," the *method of residues,* expresses more clearly than the others the eliminative function of all the canons. Its statement is: *"Subduct from any phenomenon such part as is known by previous inductions to be the effect of certain antecedents, and the residue of the phenomenon is the effect of the remaining antecedents."* [10]

The method very clearly depends upon our making use of some already known causal connections in order to isolate the influence of some other known or assumed cause by means of a *strictly deductive argument.*

A favorite illustration for this method is the discovery of the planet Neptune by Adams and Le Verrier. The motions of the planet Uranus had been studied by the help of Newton's theories. Its orbit was plotted on the assumption that the sun and the planets within the orbit of Uranus were the only bodies which determined its motion. But the calculated positions of Uranus were not in agreement with the observed positions. On the assumption that these differences could be explained by the gravitational action of a planet outside the orbit of Uranus, the position of such a hypothetical planet (behaving according to the usual principles of celestial mechanics) was calculated from the perturbations in the motion of Uranus. And in fact the planet Neptune was discovered in the vicinity of the place calculated for it. This achievement is therefore credited to the method of residues.

---

[10] *A System of Logic,* Vol. I, p. 460.

But the argument used for locating Neptune is easily seen to be strictly deductive. We must accept, in the first place, the universality of Newton's theory of gravitation. We must assume, in the second place, that the motion of Uranus is determined by the known bodies within its orbit and a single unknown body outside its orbit. The position of this unknown body can then be calculated if we also know how much of the observed behavior of Uranus is due to the influence of the interior planets. Now the canon of residues itself did not pick out the cause of the discrepancies in the observed behavior of Uranus. A hypothesis concerning the possible source of this discrepancy had to be explicitly introduced. The canon simply expresses the fact that, on the assumptions made, the interior masses were to be eliminated as the causes of the observed discrepancies. It does not suggest *where* the source of the residual phenomena is to be located. It does not demonstrate that the suspected source of such residual phenomena is causally related to them.

In this illustration one further condition for the applicability of this method must be noted. We can calculate the position of the planet Neptune only if we know the law according to which forces of attraction can be compounded. These forces are supposed to act "independently" of one another. This means that if one of the interior planets should fly off from the solar system, the magnitude of acceleration that each of the remaining bodies would contribute to the behavior of Uranus could still be calculated from their known positions and masses. Wherever the forces studied are not independent in this sense (where, in other words, the effect of two forces cannot be calculated from a knowledge of each in isolation) the method of residues cannot be employed.

§ 8. SUMMARY STATEMENT OF THE VALUE OF THE EXPERIMENTAL METHODS

We will now summarize this long discussion of the experimental canons. Every investigation of the cause of a phenomenon *P* must start with a hypothesis. Suppose $H_1, H_2, \ldots H_n$ are a set of alternative hypotheses concerning the possible determining conditions of *P*. The *H*'s therefore express our sense of what is relevant in any occurrence of *P*. No observation, experiment, or reasoning can proceed without an explicit or implicit acceptance of the following:

Proposition 1. *Either* H_1, *or* H_2 *or . . .* H_n *is the causal law of* P. The function of the experimental canons is to eliminate some or all of these alternatives. We try to show that in the instances where P is present, H_1 does not hold; or that H_1 is true in those cases where P is absent as well as where P is present; or that a variation occurs in P without a correlated change in some factor denoted by H_1. If we are successful in showing any one of these things (and if H_1 represents a proper analysis of circumstances), H_1 is then eliminated as a causal law of P. Thus experiment may establish the following:

Proposition 2. H_1 *is not the causal law of* P.

We may then conclude from propositions 1 and 2:

Proposition 3. H_2 *or* H_3 *or, . . .* H_n *is the causal law of* P.

The same procedure can now be undertaken for H_2, and so on. And we may be successful in eliminating all the alternatives but H_n. Provided that H_n cannot be eliminated, we may conclude that H_n is the causal law of P, on the assumption that the nH's are the only possible causal laws.

But it is clear that this procedure is efficacious in finding causal laws only if the following be true.

a. Proposition 1 must be based upon a proper analysis of the circumstances attendant upon P. The H's must express the relevant relations of P to certain other factors.

b. The n alternatives H must include the true causal law of P. If we have not been fortunate enough to include the true causal law, all the alternatives may be eliminated and the cause of P not be ascertained. But no directions can be given how to include the true law in an enumeration of possible laws. The difficult step in extending our knowledge consists, therefore, in finding propositions of the form *If* H *then* P where H is a suitable hypothesis or theory from which the phenomenon P can be shown to be a consequence.

c. Proposition 3 is obtained by strictly necessary reasoning from propositions 1 and 2.

d. The concluding propositions are not *demonstrated* to be true unless propositions 1 and 2 are in fact true. But we can rarely, if ever, be certain that proposition 1 is an exhaustive statement of all possible causal laws for P.

The canons of experimental inquiry are not therefore capable of *demonstrating* any causal laws.

The experimental methods are neither methods of proof nor methods of discovery. The canons which formulate them state in a more explicit manner what it is we generally *understand* by a causal or invariant relation. They *define* what we mean by the relation of cause and effect, but do not *find* cases of such a relation. The hope of discovering a method that will "leave little to the sharpness and strength of men's wits" is one which finds no support from a careful study of the procedure of the sciences.

But while the methods we have discussed have the defects pointed out, they are of undoubted value in the process of attaining truth. For in eliminating false hypotheses, they narrow the field within which true ones may be found. And even where these methods may fail to eliminate all irrelevant circumstances, they enable us with some degree of approximation to so establish the conditions for the occurrence of a phenomenon, that we can say one hypothesis is logically preferable to its rivals.

§ 9. THE DOCTRINE OF THE UNIFORMITY OF NATURE

The claim that the experimental methods are capable of demonstrating with complete certainty universal, invariable connections rests on a belief that "nature is uniform." Induction, according to Mill, consists in inferring from a finite number of observed instances of a phenomenon, that it occurs in *all* instances of a certain class which resemble the observed instance in certain ways. But according to Mill, the very statement of what induction is requires an assumption concerning the order of the universe. The assumption is that "there are such things in nature as parallel cases, that what happens once, will, under a sufficient degree of similarity of circumstances, happen again." [11]

This assumption may be expressed in various ways: that nature is uniform, that the universe is governed by general laws, that the same cause will under similar circumstances be accompanied by the same effect. In *some* form, however, so the claim runs, it is required for induction. Every induction may be thrown into the form of a syllogism, and the principle of the uniformity of nature will then appear as the "ultimate major premise of all inductions." [12]

Mill puts the matter as follows: "The induction, 'John, Peter, etc., are mortal, therefore all mankind are mortal,' may . . . be thrown into a syllogism by prefixing as a major premise (what is at any rate a necessary condition of the validity of the argument) namely, that what is true of John, Peter, etc., is true of all mankind. But how came we by this major premise? It is not self-evident;

---

[11] *A System of Logic,* Vol. I, p. 354. [12] *Ibid.,* p. 356.

nay, in all cases of unwarranted generalization, it is not true. How, then, is it arrived at? Necessarily either by induction or ratiocination; and if by induction, the process, like all other inductive arguments, may be thrown into the form of a syllogism. This previous syllogism it is, therefore, necessary to construct. There is, in the long run, only one possible construction. The real proof that what is true of John, Peter, etc., is true of all mankind, can only be that a different supposition would be inconsistent with the uniformity which we know to exist in the course of nature. Whether there would be this inconsistency or not, may be a matter of long and delicate inquiry; but unless there would, we have no sufficient ground for the major of the inductive syllogisms. It hence appears, that if we throw the whole course of any inductive argument into a series of syllogisms, we shall arrive by more or fewer steps at an ultimate syllogism, which will have for its major premise the principle, or axiom, of the uniformity of the course of nature."[13]

We shall not discuss whether the principle of the uniformity of nature is true or whether some such principle is required for making inductive inferences. We wish simply to determine whether the principle *if it were true* would in fact help to demonstrate the existence of some particular instance of a supposed causal relation. We must carefully note the following.

1. The principle is stated in an extremely vague form—"what happens once, will, under a sufficient degree of similarity of circumstances, happen again." But what is a sufficient degree of similarity? The principle does not tell us. In any particular investigation we must rely on other criteria, if there are any, to determine what are the circumstances material to the occurrence of a phenomenon.

2. In the second place, the minor premise of an inductive syllogism, even according to Mill, is a *particular* proposition. Therefore even if we employ a universal major premise, such as the principle of uniformity of nature, the premises are insufficient to *demonstrate* a universal conclusion.

3. Finally, the principle does not affirm that *every* pair of phenomena are invariably related. It simply states that *some* pairs are so connected. To appeal to the doctrine in a particular investigation is therefore useless. If we suspect that tight-fitting hats are the cause of baldness, we employ the canons to eliminate as many circumstances other than tight hats as we can. But no finite number

---

[13] *A System of Logic,* Vol. I, pp. 357-58.

of observed cases of tight-fitting hats followed by baldness can demonstrate a law which is to hold for an *indefinite* number of cases. The principle of uniformity of nature does not help us. It does not say *which* of the innumerable casual connections between phenomena are invariable; it merely asserts that *some* are. But the task of the particular inquiry is to show that a *designated* pair of phenomena are in causal relation.

§ 10. THE PLURALITY OF CAUSES

The method of agreement is often found to be faulty because we cannot be sure—so it is claimed—that the effect studied may not have more than one cause. It is for this reason that the method of difference was regarded by Mill as a superior experimental procedure. The doctrine of plurality of causes is stated by him as follows: "It is not true . . . that one effect must be connected with only one cause, or assemblage of conditions; that each phenomenon can be produced only in one way. There are often several independent modes in which the same phenomenon could have originated. One fact may be the consequent in several invariable sequences; it may follow, with equal uniformity, any one of several antecedents, or collection of antecedents. Many causes may produce mechanical motion: many causes may produce some kinds of sensations: many causes may produce death. A given effect may really be produced by a certain cause, and yet be perfectly capable of being produced without it."[14]

This doctrine may be given a logical version. The fallacy of affirming the consequent in mixed hypothetical syllogisms may be interpreted as an illustration of the doctrine of plurality of causes. Thus, given *If a number expressed in ordinary algorism has a 5 in the unit place, it is divisible by 5,* we cannot validly infer that a certain number terminates with a 5 in the unit place because it is divisible by 5; the number *may* terminate with a zero. It seems, therefore, that Mill is in the right concerning plurality of causes, and his doctrine is capable of being stated in a more general form and in purely logical terms.

Let us first consider the less general form of the doctrine, as stated by Mill. Suppose a house burns down. What is the cause of this event? Perhaps the house was destroyed because of an overturned kerosene lamp, or because of defective electric wiring, or because of a faulty chimney. The reader may be tempted to retort that the plurality is only apparent. "If the alleged causes of the fire were examined more carefully," he may perhaps say, "a circumstance common to all of them would be found. For example, the occurrence of a rapid oxidation in some part of the house is such a common circumstance. And that common feature of the many alleged causes is *the* cause of the event."

---

[14] *Ibid.*, p. 505.

Such an analysis is not very satisfactory. If the reader were investigating that fire for an insurance company and submitted such an analysis, he would not retain his post for long "The occurrence of a rapid oxidation," the company would doubtless declare, "is an explanation of all fires. It was not your job to discover the most general conditions under which fires occur, for we knew that all the time; it was your job to find the special conditions under which this one occurred."

This hypothetical reply of the insurance company not only indicates the inadequacy of one type of criticism of the doctrine of plurality of causes; it also suggests a more complete reply to the doctrine. For if the doctrine were true, how could we ever be able to infer the cause from an examination of the ruins of a destroyed house? There can be no doubt that we frequently infer the true cause of an effect. Fire insurance companies do so continually. So also the medical examiner is able to establish the real cause of a person's death in spite of the alleged plurality of causes of death.

The more satisfactory reply to the doctrine of plurality of causes is this: When a plurality of causes is asserted for an effect, the *effect* is not analyzed very carefully. Instances which have significant differences are taken to illustrate the *same effect*. These differences escape the untrained eye, although they are noticed by the expert. Thus the way in which a house burns down when an overturned lamp is the cause, is not the same as when defective wiring is the cause. The doctrine of plurality of causes is plausible only if we analyze the causes into a larger number of distinct types than we do the effect. The doctrine overlooks many differentiating factors present in several instances of a so-called effect, and by viewing these instances under their more generic features regards them as instances of the *same* effect. For many purposes it is perhaps convenient to retain this lack of symmetry in the analysis of causes and effects. But it does not follow from this fact of convenience that the usual illustrations of plurality of causes really prove the absence of a one-to-one correspondence between cause and effect.

Let us now turn to the doctrine in its more general or logical form. Must we deny that the fallacy of affirming the consequent is a fallacy? Not if we recognize elementary distinctions, and recall some of our discussions in the chapter on mathematics. Affirming the consequent is a fallacy because the same consequent may follow from more than one antecedent. But, we may ask, if a proposition follows from two distinct sets of premises, does it follow from them in virtue of their being different from each other, or in virtue of their containing something in common?

If the reader remembers our discussion of logical systems, he must acknowledge that the second alternative expresses the true state of affairs. We showed in Chapter VII, § 3, that two systems may be incompatible with each other taken in their entirety although they may have many theorems in common. We explained this by suggesting that the two systems contain a common subsystem. The common theorems of the two systems follow strictly from the axioms of this common subsystem, and not from the axioms of the two systems as such.

Let us state this in different words. A set of premises which is a sufficient condition for a given proposition may contain conditions besides those which are necessary. With care and ingenuity, those portions of the premises which are not required for the conclusion can be eliminated. In this way, we can discover all the conditions necessary for the conclusion. And when the antecedent in an implicative proposition contains the necessary and sufficient conditions for the consequent, it is no longer a fallacy to argue from the affirmation of the consequent to the affirmation of the antecedent.

The fallacy of affirming the consequent is therefore indeed a fallacy, since we do not in general know that the antecedent states the necessary and sufficient conditions for the consequent. For most purposes, science is satisfied with the sufficient conditions for the propositions it wishes to establish. But its goal, which may never be reached, is to find the conditions which are both necessary and sufficient.

The distinction between sufficient conditions and those which are both necessary and sufficient throws further light on the limitations of so-called crucial experiments. Suppose *p*, a verifiable proposition, follows from theory T_1 but not from T_2. Then a slight modification of T_2, one which leaves its main outlines unaffected, may perhaps be made so that *p* is implied by the revised T_2 as well as by T_1. Both theories, in such a case, would logically contain the necessary and sufficient conditions for *p*, although they may contain much else besides. The verification of *p*, therefore, will not compel us to abandon T_2 if we can continue to use it with scientific profit after slightly altering it. We may conclude with a pertinent remark of Bertrand Russell: "a hypothesis which accounts with a minute exactitude for all known relevant facts must not be regarded as certainly true, since it is probably only some highly abstract aspect of the hypothesis that is logically necessary in the deductions which we make from it to observable phenomena." [13]

---

[13] *The Scientific Outlook,* 1931, p. 67.

CHAPTER SIX

CLASS NOTES

We have omitted class notes for this chapter, since the student will already have had a fair number of exercises in dealing with measures of association in Chapter Two. Further exercises involving measures of association, found in the class notes for Chapter Seven, introduce the problem of dealing with a third variable.

CHAPTER SEVEN

ANALYSIS OF DATA

I

It is a commonplace in the philosophy of science that facts never "speak for themselves." Before facts are to have any meaning, they must be disentangled and given order. They must be weighed and fitted to a relevant body of ideas or theories.

We have included three readings in this chapter. The first, "Split Trials and Time Saving" by Hans Zeisel and Thomas Callahan, represents, we think, an excellent example of a project joining the law and the methods of research often found in the social sciences. The data are analysed in a relatively simple, clear-cut manner with a relatively simple, specific question in mind.

The second selection, "Some Functions of Qualitative Analysis in Social Research" by Allen H. Barton and Paul F. Lazarsfeld, ranges more broadly. Fundamentally, the article is concerned with the act of imagination, the process of trying "to make sense" out of the infinite complexities which lie concealed in the data dredged up by empirical inquiry. And the authors of the article implicitly recognize that in some sense this process must be akin to artistic creation, just as in the case of formulating the hypotheses with which we begin our research. This is not to say that the "truth" revealed by the analysis of our data is arbitrary or subjective or free from the canons of scientific proof. It simply means that there is no one correct way to analyse the data, no mechanical set of procedures which must be followed in every instance.

The last reading is by William J. Goode and Paul K. Hatt and it describes some methods of analysing and presenting data, such as cross-tabulations, frequency, distributions, and so on, which are already familiar. In addition, however, the authors touch on a new and much more complex issue--namely the introduction of a third factor into the analysis of relationships. As the reader will see, Goode and Hatt employ terms such as interpretation, explanation, and specification, which are part of the language of the social sciences, to distinguish different causal interpretations. But the basic idea is fundamentally the same: The relationship between two characteristics is frequently modified by the influence of a third characteristic; and this influence must be statistically controlled or held constant, if we are to increase our understanding.

I I

With these remarks in mind, we can list some of the modes of data analysis which are frequently used. They are very basic ways of looking at data which crop up again and again when we examine the results of research projects. Much of the following material has already been covered, of course, for the analysis of data involves such things as examining the procedures of sampling, the simplification of data, tests of significance, and so on. We are listing these aspects of research here, however, not as steps to be followed in the collection of data but as items to be taken into account when we ask what our data mean and how far we can generalize our results.

1. If we draw a random sample from a population of people or of events or of some other type of unit, we can make generalizations about the population from which it comes, as we have seen. But can we make generalizations to other similar populations? What can one sampled community, for example, tell us about other communities? Will the conclusions about a population, based on a sample, be valid at other points in time? There is no definite answer to these questions--"it all depends," we say--but they are not questions that can be avoided in the analysis of data.

It is a very common practice to extend the conclusions of a study beyond the population which has been examined. Either implicitly or explicitly, it is frequently argued that the conclusions drawn from a study do have implications that go beyond the immediate case in hand. But this is not an assertion that can be made blindly--we must note the similarities and differences of the population studied in comparison with the larger set of populations which concern us and come to a reasoned judgment of how these similarities and differences will influence our generalizations.

Still, we must make sure our starting point is a strong one--that is, we must be as certain as we can that one sample studied is representative of its universe. One of the first things to do, then, in analysing data, is to examine carefully and critically the sampling procedure used.

2. Even a random sample, however, may not be representative. This is a point to be checked, if at all possible, by comparing the characteristics of the sample with whatever information is available about the universe from which it comes. Census materials, official records, previous surveys--all may provide summary information about the population which can be used to gauge how close the particular sample of a particular study comes to mirroring the parent population.

3. In the material on the collection of data, it was pointed out that there are many steps between empirical reality and the records that end up in research project files. When we examined the so-called instru-

ments used by the social scientist to collect data--interview schedules, unstructured questionnaires, protocols for recording non-verbal behavior, etc. --we must remember to ask if alternative methods of data collection would produce the same results. Even with one particular method, such as an interview schedule used in a survey, important changes in the end product can occur with changes in the wording of questions, the order in which they are asked, and so on.

4. We have continually stressed the concept of relationship in these materials, arguing that the concept was the very stuff of which scientific reasoning is made. One of the most important types of relationships involves the covariation of two variables, for it provides a model for much of our thinking about relationships between characteristics in the form of attributes. When, however, we are actually dealing with variables rather than characteristics such as attributes, in looking at the data accumulated in a particular project, we must be very sure that certain assumptions underlying our calculations are valid. One of the most important of these is the difference between rectilinear and curvilinear relationships discussed in some detail in the readings in the chapter on correlation. If we make a mistake on this point, we can seriously distort the reality embedded in our data.

5. As indicated earlier, we often find it convenient to simplify our data when we wish to get at the underlying patterns of relationships. Sometimes this is a matter of calculating means and standard deviations, percentages, and so on. However, it may also involve converting a variable or a set of ranked categories into a dichotomous attribute. The problem here is one of deciding on the cutting point--and although there are no hard-and-fast rules on this issue, we often try to break the frequency distribution approximately in half. The dichotomizing of variables into attributes means that we lose a good deal of detail in our data; but it also may mean that the analysis is a good deal less complex.

6. We often find that characterizing a unit of observation in terms of one characteristic at a time is not as revealing as we would like it to be. We need to construct scales, indices, and other forms of scoring devices, which are really ways to characterize units of observation with a great deal of precision along a single dimension. This issue is beyond the scope of our introduction to research methods in the social sciences, but the student should at least be familiar with one elementary form of dealing with multiple characteristics, namely, the construction of typologies based on two dichotomized attributes. We may, for example, in analysing data, classify individuals into "high" or "low" in terms of two characteristics; and from the resulting 2 x 2 table we can construct a new classification of

individuals--a typology--based on the joint occurrence of the attributes. Thus, Type I would be "High-High"; Type II "High-Low"; Type III "Low-High"; and Type IV "Low-Low." The usefulness of this simple procedure is well presented in the reading by Barton and Lazarsfeld.

7. Once we have established the characteristics we wish to examine (variables, attributes, types, etc.) our next task is to examine their inter-relationships. An important part of this task involves checking the nature of this relationship, i.e., is it curvilinear or rectilinear? In addition, however, we are concerned with the extent of the relationship--and for this we need some form of measurement. The readings have touched on two forms of measurement which are frequently encountered in the analysis of data: Yule's Q for attributes and Pearsonian Product-Moment r for variables. The student should be familiar with the interpretation of these measures, since he will inevitably encounter them in the literature of the social sciences.

8. The need to sample a population, rather than examining every unit it contains, has been argued before at some length. And, as we have pointed out, the sample gives our best estimate about the population or universe from which it comes. We are still faced with the task of judging the accuracy of our estimate, however, and this involves establishing confidence limits and making tests for the significance of differences. The

important point to remember, in the analysis of data, is that as valuable as a sample may be, it cannot be taken at face value. We must logically work out the limits of its accuracy.

9. The examination of relationships may indeed by the stuff of which science is made, as we have argued, but it is clear that the complexities of the empirical world demand more than the examination of two characteristics at a time. The task of ascertaining the influence of other characteristics can grow extremely difficult and complex, but we can limit ourselves to three forms which are fundamental.

First, we can ask if an observed relationship can be accounted for by an intervening variable, in the sense that the independent variable causes the intervening variable (Goode and Hatt refer to the intervening variable as the test variable) which in turn causes the dependent variable. If we can show (1) that the independent variable is related to the intervening variable; (2) that the intervening variable is related to the dependent variable; and (3) that the relationship between the independent and dependent variable is much diminished or disappears when the intervening variable is held constant, we have provided evidence for our theory. (Goode and Hatt give us an illustration with attendance at a private or public secondary school as the independent variable, secondary school grades as the intervening variable, and grades in college as the dependent variable.)

Second, we can ask if an observed relationship can be accounted for by an antecedent variable, in the sense that two characteristics are related not because one causes the other, directly or indirectly, but because both are caused by a third variable--the so-called antecedent variable. The statistical manipulation is the same as in the case where we try to determine the influence of an intervening variable--we hold the third variable, i.e., the antecedent variable, constant and see if the observed relationship disappears. (Goode and Hatt use economic status and birth rate as independent and dependent variables respectively, and education as a possible antecedent variable.)

Third, we can analyse our data to see if an observed relationship between two characteristics increases, decreases, or remains the same when a third characteristic is held constant. That is to say, we measure the observed relationship for all those cases in which the third characteristic is present and we measure the same relationship for all those cases in which the third characteristic is absent. This examination of the relationship under varying conditions is referred to as specification; and Goode and Hatt use as an illustration the relationship between education and rank in the Army under conditions of long-term or short-term service. Again, the construction of the necessary tables is the same as in the two previous types of analysis, although there is a difference in the nature of the theoretical issues which are being demonstrated.

III

With this chapter on the analysis of data, we come to the end of our introduction to the union of law and social science research. We have tried to stress what we think are the most fundamental ideas and we have tried to avoid a too narrow concern with specific techniques. But these materials are not viewed as some definitive set of readings which must be covered in a rigid and exhaustive fashion. Teaching materials such as these, we believe, must be used in a flexible manner, with the teacher and students adapting them to their particular interests and background.

In our experience with using these materials in the law school, we have found that law students are usually more concerned with the substantive issues--at least initially--than with the methodological issues. This is understandable enough; and, in fact, we think that teaching social science methodology to law students successfully requires a continuous emphasis on substantive application. At the same time, we think it is worthwhile to reiterate the basic argument on which this collection of readings is founded--namely, that the judicious use of substantive material drawn from the social sciences is dependent on at least an elementary understanding of the research methods by which such material is assembled. If this collection of readings aids the law student in achieving that understanding, we will have achieved our purpose.

SIMPLE TECHNIQUES AND IMPORTANT RESULTS

Hans Zeisel and Thomas Callahan were asked by the District Court for the Northern District of Illinois to study the time saved by splitting trials into two separate proceedings--one to determine liability and one to determine damages. We think that their research provides an excellent example of the practical value in social science in solving pressing legal problems. There are two points particularly worth noting: (1) the research design is an approximation of a true experiment; and, (2) the authors use relatively simple mathematics and yet produce important results.

In discussing random sampling procedures, Mueller and Schuessler in Chapter Four make the point that the "ideal sampling procedure is one in which the drawings are affected by impartial chance factors alone." This, of course, is a device to insure that the samples present an accurate representation of the universe under observation. Zeisel and Callahan, due to the technicalities surrounding the adoption of a judicial rule, could not use random selection procedures. Rather, they had to rely upon data "routinely collected by the Office of United States Courts." As a result, we cannot be sure if the cases tried by one method are the same as those cases tried by another--and, as a consequence, we cannot be sure if any difference in outcome is due to the difference in the method of trial or some other factor. The student should note how the authors come to grip with this problem.

As for the use of simple mathematics, the reading makes it clear that the law student can engage in important research without necessarily using formulas for correlation, tests for significance, etc. In fact, the authors rely heavily on simple percentages and simple logic--but the methods fit their purpose and the results are more than adequate.

From Hans Zeisel and Thomas Callahan, "Split Trials and Time Saving: A Statistical Analysis," Harvard Law Review, 76 (1963), pp. 1606-1625. Reprinted by permission.

6

Hans Zeisel* and Thomas Callahan**,
"Split Trials and Time Saving: A
Statistical Analysis," Harvard Law
Review, 76 (1963), pp 1606-1625.

In 1959 the United States District Court for the Northern District of Illinois adopted a rule permitting separate trials of liability and damage issues in civil cases. At the request of the court the authors have conducted an investigation to determine the extent to which time is saved through use of the separation device. They use a variety of converging statistical approaches to arrive at the conclusion that in personal injury jury trial cases about twenty per cent of trial time may be saved. They then consider the possible ways that separation might lengthen trial time, and conclude that these will not offset the time saving to any degree.

Among the devices aimed at reducing the trial load of our courts, the split or separate trial of issues has been proposed as holding special promise. Trial time would be saved, the argument goes, if the issues of liability and damages were tried separately. The jury would be required first to bring in a verdict on liability, and only if liability were found would the jury hear evidence and render a second verdict on the issue of damages.[1] On the recommendation of the late Judge Julius H. Miner,[2] the United States District Court for the Northern District of Illinois was persuaded to test this device, and on November 3, 1959, adopted civil rule 21, set forth below,[3] permitting separate trial on motion of a party

---

† We are happy to acknowledge the generous support of this study by the Walter E. Meyer Research Institute of Law. This investigation, an offshoot of the University of Chicago Law School's prior work on the jury system and court congestion, was conducted in closest collaboration with Professor Harry Kalven, Jr., Director of the Jury Project.

* Professor of Law and Sociology, University of Chicago Law School. Dr. jur., University of Vienna, 1927, Dr. rer. pol., 1928.

** Third-year student, University of Chicago Law School; Research Assistant, Jury Project, since 1957. M.A., University of Chicago, 1962.

[1] Such a plan was proposed in ZEISEL, KALVEN & BUCHHOLZ, DELAY IN THE COURT 99 (1959) [hereinafter cited as DELAY].

[2] Miner, *Court Congestion: A New Approach*, 45 A.B.A.J. 1265, 1268 (1959). Judge Miner died on March 11, 1963, the very day this report was submitted to his court. The cause of court administration, and especially its students, owe him a great debt of gratitude for his determined efforts to pioneer what promises to be a uniquely powerful delay remedy.

[3] N.D. ILL. (CIV.) R. 21, set forth in 2 FED. RULES SERV. 2D 1048-49 (1960):

> Pursuant to and in furtherance of Rule 42(b), Federal Rules of Civil Procedure, to curtail undue delay in the administration of justice in personal injury

or at the direction of the court. Though the rule was sponsored in order to save trial time, a commodity in short supply in most metropolitan courts, attempts were soon made to upset it on the ground that this major departure from the traditional mode of trial might affect the substance of the verdicts rendered by juries. But the Court of Appeals for the Seventh Circuit held that the rule preserved the essential character of trial by jury and was within the ambit of Rule 42(b) of the Federal Rules of Civil Procedure, which permits separated trial of issues in the interest of convenience.[4]

Prior to the promulgation of the rule, Chief Judge Campbell asked the Law School of the University of Chicago to design a research operation that would measure the effect of separation on the trial load of the court.[5] A later report may deal with the above-mentioned substantive considerations; the present report treats only of one aspect of the rule — its impact on trial time.

The way in which separation of issues may save court time is clear. In the traditional form of trial, the damage issue must be litigated even where the verdict will ultimately reject liability; separation would eliminate the need for trying the damage issue in those cases, comprising roughly 40 per cent of all personal injury jury trials.[6] Even where liability was affirmed, it seemed likely

---

and other civil litigation wherein the issue of liability may be adjudicated as a prerequisite to the determination of any or all other issues, in jury and non-jury cases, a separate trial may be had upon such issue of liability, upon motion of any of the parties or at the Court's direction, in any claim, cross-claim, counterclaim or third-party claim.

In the event liability is sustained, the Court may recess for pre-trial or settlement conference or proceed with the trial on any or all of the remaining issues before the Court, before the same jury or before another jury as conditions may require and the Court shall deem meet.

The Court, however, may proceed to trial upon all or any combination of issues if, in its discretion, and in furtherance of justice, it shall appear that a separate trial will work a hardship upon any of the parties or will result in protracted or costly litigation.

[4] Hosie v. Chicago & N.W. Ry., 282 F.2d 639 (7th Cir. 1960), *cert. denied*, 365 U.S. 814 (1961). The court did not decide whether the use of different juries for liability and damage issues as authorized by the district court rule violates the seventh amendment. The district court has confined the operation of rule 21 to single juries. *Cf.* O'Donnell v. Watson Bros. Transp. Co., 183 F. Supp. 577 (N.D. Ill. 1960).

[5] Throughout the study Judges Campbell, Miner, and Robson acted as advisers, for which we owe them a great debt of gratitude. We also want to thank the other members of the court and its clerks for their meticulous help throughout these two years. We are especially indebted to Mr. Roy Johnson, Special Assistant to Judge Campbell, and to Miss Aileene C. Coldby, Chief Deputy Clerk, for their untiring cooperation. Appreciation is also due to Mr. Orin S. Thiel and Mr. Ronald H. Beatty of the Administrative Office of the United States Courts for their thoughtful help in supplying us with a series of control statistics from their office.

[6] From a nationwide sample of civil jury trials, which forms part of the basic data of the Jury Project of the University of Chicago Law School.

that trial time might be saved by the stimulus to settlement before the damage issue is litigated. There remained, however, the possibility that these savings might well be offset by a number of countervailing factors not so immediately obvious. Thus the separation of issues might reduce the proportion of cases settled, and thereby increase the number of cases requiring trial. It might reduce the ratio of jury waivers and thereby add to the number of cases requiring jury trial.[7] Separation might increase the proportion of hung juries, and necessitate more retrials. Finally, it might simply require more time of juries overall, since in some cases there would be two deliberations. The crucial question for this study was, therefore, not so much whether but rather *how much* time would be saved by the separation of issues.

Because our inquiry was set up only shortly before the adoption of the separation rule, we were unable to acquire data on the period before the rule which we should have wished to have in order to compare the operations of the court before and after the separation rule was introduced. Instead, for our prior control period we had to rely for information on data routinely collected by the statistical branch of the Administrative Office of the United States Courts. The ideal research design would have combined such a prior observation period with a selection by chance of the cases to which the separation rule would be applied. Specifically, one might have subjected every three successive cases filed to a sort of lottery: in the case drawn first, separation would be required; in the second, separation would not be allowed; in the third case, separation would be left to the judge's discretion. These groups of three would be assigned in rotation, giving each judge an equal number of cases of each variant. After final disposition, the separated, the regular, and the cases where separation was discretionary would be compared and measured against the period that preceded the introduction of the rule. But this division of cases was of course not feasible for a variety of good reasons.[8] Instead, all that could be done was to keep a careful record of the court's actions in each case from assignment to final disposition, both for cases in which there was separation and for the regular trials. We had no influence and of course did not attempt to exert any influence on the choice of cases in which the issues were separated. Each judge applied the broad discretion conferred on him by the rule according to his own understanding of its suitability for the case at hand.

---

[7] The New York experience showed that roughly 40% less time is used up in a bench trial. DELAY 81.

[8] *Cf. The Case for the Official Experiment*, DELAY 241–50; Zeisel, *The New York Expert Testimony Project: Some Reflections on Legal Experiments*, 8 STAN. L. REV. 730, 741–48 (1956).

I

To begin the analysis, then, we present a synopsis of the civil cases tried before the court. Table 1 gives the proportion of cases tried under the separation rule in each category. The table shows that the relative frequency of such trials varies greatly with the type of action. Two categories show no separation whatsoever: admiralty cases and tort cases other than personal injury claims. The greatest use of separation was made in personal injury trials: here 37 per cent of the jury cases and 27 per cent of the bench trials were tried under the separation rule. Separation was ordered

TABLE 1

RELATIVE FREQUENCY OF SEPARATED TRIALS BY TYPE OF CASE [a]

		Regular Trials	Separated Trials			
		Number	Number	Percentage of Total	TOTAL	
Personal Injury: [b]	Jury	117	69	37%	186	
	Nonjury	22	8	27%	30	216
Other Tort:	Jury	10	..	..	10	
	Nonjury	4	..	..	4	14
Contract:	Jury	19	6	24%	25	
	Nonjury	17	2	11%	19	44
Federal Tort Claims Act:	Nonjury	10	4	29%	14	14
Admiralty:	Jury	1	..	..	1	
	Nonjury	6	..	..	6	7
Other:	Jury	8	..	..	8	
	Nonjury	34	4	11%	38	46
TOTAL:	Jury	155	75	33%	230	
	Nonjury	93	18	16%	111	
		248	93	27%	341	

[a] For the two-year period after introduction of rule 21 (1960–1961).

[b] Includes 16 jury and 6 nonjury FELA cases, of which one jury trial was separated.

in only one out of twenty-two FELA cases but in 29 per cent of the fourteen Federal Tort Claims Act cases. These frequencies forced us to limit our inquiry to personal injury cases: only this category provided a sufficiently large group of separated (69) and regular (117) trials to offer an adequate basis for statistical comparison. It would be improper to compare, for instance, separated tort cases with nonseparated contract cases, since the validity of any comparison requires that the cases be kept as homogeneous as possible. Strictly speaking, therefore, conclusions of the study will apply only to personal injury jury cases, but within reason they should prove applicable also to other types of cases.

Table 2 proceeds to the comparison of separated and regular personal injury trials. The 186 trials by jury during the observation period are divided by the stage at which they terminated. We distinguish four stages for the separated trials and four roughly parallel stages for the regular trials. The midpoint for the regular trial comes at the end of the plaintiff's case; for the separated trial, at the end of the liability trial.

TABLE 2

Stages of Disposition and Time Required To Reach Them[a]

Separated Trials			Regular Trials		
Stage of Disposition	Days	Per Cent	Stage of Disposition	Days	Per Cent
During Liability Trial	1.9	19%	During Plaintiff's Case	2.0	18%
At End of Liability Trial	3.2	62%	At End of Plaintiff's Case	3.3	4%
During Damage Trial	3.6	4%	During Defendant's Case	..	..
After Full Trial	4.0	15%	After Full Trial	4.7	78%
Average All Cases	**3.1**	100%	Average All Cases	**4.2**	100%
Number of Cases	(69)		Number of Cases	(117)	

[a] Personal injury jury trials. Only court time is considered; deliberation time is not included. Figures in the "Days" column are averages for cases disposed of at the stage indicated.

It was, to repeat, the expectation of those who favored the new rule that many separated trials would end after the liability verdict, thus making the damage trial unnecessary. As can be seen, Table 2 vividly confirms this expectation: fully 78 per cent of the regular trials, but only 15 per cent of the separated trials, go through their full course. And since many more regular trials run their full course than do separated trials, it must be expected that the average trial time for each is quite different. The figures for trial time in Table 2 bear this out, showing that the average regular trial lasts 4.2 days, the separated trial, 3.1 days — a time difference of 26 per cent. It would be tempting to read this figure as the looked-for measure of the time saved by separation. Unfortunately, it cannot be accepted at face value because, standing by itself, this comparison of the separated and the regular trials could be quite misleading.

II

At this point we ask for the reader's indulgence during a somewhat complicated methodological excursion that will, we trust, make clear why without further investigation we cannot trust the comparison made by Table 2. The precise question is this: Were the cases selected for separated trial and the cases left for regular trial substantially similar except for the fact of separation? If they were, we may attribute the observed difference to the fact of separation. But if they were not, we are in trouble. It is the same methodological problem that came to the fore in the now famous debate over whether smoking shortens our life expectancy. The average age at which smokers die is lower than that of nonsmokers, just as the average separated trial is shorter than the average nonseparated one. Does such a statistic prove that smoking shortens life, or that separation shortens trial time? Suppose it is the man who is less healthy to begin with who takes up smoking. We would then falsely attribute to smoking the effect of shortening our lifetime, when in fact the average smoker had a shorter life expectancy even before he began smoking. Similarly, one may find that deaths are more likely to occur in the near future in families recently visited by physicians than in those unattended. Obviously it is not the doctor, but the preexisting situation that calls for a doctor, which causes the higher mortality rate. A similar difficulty may apply to the figures in Table 2: cases selected for separate trial might be different to begin with. If that difference is itself related to the time saving we want measured, then we no longer know to what extent the apparent time saving was due to differences other than separation. That was why our ideal research design specified selection by lot — by avoiding purposeful selections we could avoid the possibility of selections on some basis which might have a bearing on trial time.

Were our separated trials selected in any way that could lessen the validity of the comparison in Table 2? We know that there was purposeful selection by the judges. Though we do not know the basis for their selection, a few statistics give us warning. From Table 2, we see that the average separated trial which actually went to the jury on the damage question took 4.0 days; but it took considerably longer — 4.7 days — to try the regular cases that ran their full course.[9] There is thus some indication, though no proof, that the cases differ by selection. There is no proof, because the figure for separated trials is based on only a small fraction of cases that go the full course; the potentially longer cases might well have been disposed of during trial. On this view the difference would disappear if all trials were full trials.

Further, the figures in Table 3 suggest that the two groups of cases were not selected at random. Only 34 per cent of the regular trials ended in favor of the defendant, in contrast to 56 per cent of the separated trials — a figure far above the average.[10] Note too that the proportion of directed verdicts is larger in the separated group. Finally, there is evidence that the regular trials left over after some cases have been removed for separated trials differ as a group from the run of regular trials before civil rule 21, when no cases were separated. Statistics supplied us by the statistical branch of the Administrative Office of the United States Courts indicate that of the 196 personal injury jury trials conducted in this court during the two years prior to the introduction

TABLE 3

OUTCOME AND MODE OF TERMINATION OF SEPARATED AND REGULAR TRIALS [a]

Mode of Termination	Separated Trials Per Cent		Regular Trials Per Cent	
Settlement	32		24	
Jury Verdict for Plaintiff	12		42	
Total in Favor of Plaintiff		**44**		**66**
Jury Verdict for Defendant	43		31	
Directed Verdict for Defendant	13		3	
Total in Favor of Defendant		**56**		**34**
Number of Trials	(69)		(117)	

[a] Personal injury jury trials.

---

[9] The trial times for the regular cases are somewhat high compared with statistics available from other courts, but there are no grounds for suspecting that this should restrict the validity of our findings. We are concerned only with the relative trial time of separated and regular trials, and this relation should not be affected by the generally longer trial time.

[10] See note 6 *supra*.

of rule 21, only 62 per cent of the cases went to a jury verdict, as against 73 per cent[11] of the regular trials after rule 21. This is another suggestion that judges might have been more likely to order separate trial for those cases which held a promise of not going through a full trial.

III

So much for the potential infirmities of the data in Table 2. Do our data permit us to move the analysis to more secure ground? Table 4 will show the proportion of personal injury jury trials tried by each judge under the separation rule: the ratio varies from 89 per cent for Judge *A* down to zero for Judges *K* and *L*. Though this varying use seems at first blush unfortunate, it is this very variation which will help to solve the problem. For clarification, we return to the smoking analogy. Just as we could not dictate when judges were to use split-trial procedure, we could not select youngsters at random, ordering one group to smoke and another not to. But we can expose one random group of smokers to "stop-smoking" propaganda and keep another random group of smokers as a control group, unexposed to this propaganda. We could then compare the longevity of the "exposed" group with that of the control group. Any difference could be attributed to the "stop-smoking" campaign.

Let us try to see more precisely how this analytical procedure bypasses the difficulty that the decision to stop smoking will not be made at random, but more likely by persons who are in some way special with respect to smoking and with respect to health in general. The point is that in this design we neither know nor need to know how the self-selection works.

Let us assume the worst possible case from the point of view of selection bias: suppose the only smokers who respond to the stop-smoking campaign were those who started smoking only recently and who are at the same time particularly health conscious. Let us assume also that it is only their lengthened life span which would cause the average life span of the exposed group to increase. Even this extreme self-selection would in no way affect the validity of the proposition that reduced smoking increases life expectancy. The reason we would not be concerned with this self-selection is that we know that in the control group, too, there would be these health-conscious smokers who only recently began to smoke, in about the same proportion as they occurred in the experimental group. And we should not care if some of the control group stopped smoking on their own, without exposure to the propaganda. For it is quite sufficient to know that, whatever the control group does, the experimental group also does, in addition to the experimental group's reaction to the stop-smoking campaign.

---

[11] Table 3, second column: 42 + 31 = 73.

Now instead of two groups of smokers, let us take two groups of trials, one conducted by Judge *X* who separates some of his trials, and one by Judge *Y* who separates none. Since the court assigns cases at random, the two groups of trials will be comparable. The only differences between the groups will be that there was separation in one but not in the other, and conceivably the differing expeditiousness of Judges *X* and *Y*. If the latter factor is discounted, and if the average trial length in the group of cases where separation sometimes occurred is below that of the cases where it was not available, we are justified in ascribing this difference to the separation. One would then expect Judges *K* and *L*, who held no separate trials, to show a higher average length of all their trials than Judge *A*, who separated in 89 per cent of his trials. In reality, of course, there are not two judges, but several judges, each having a different propensity to separate. We should, therefore, expect this proposition to hold true: the larger the proportion of separated trials for the particular judge,

TABLE 4

SEPARATION RATIO AND AVERAGE TRIAL TIME IN PERSONAL INJURY JURY TRIALS

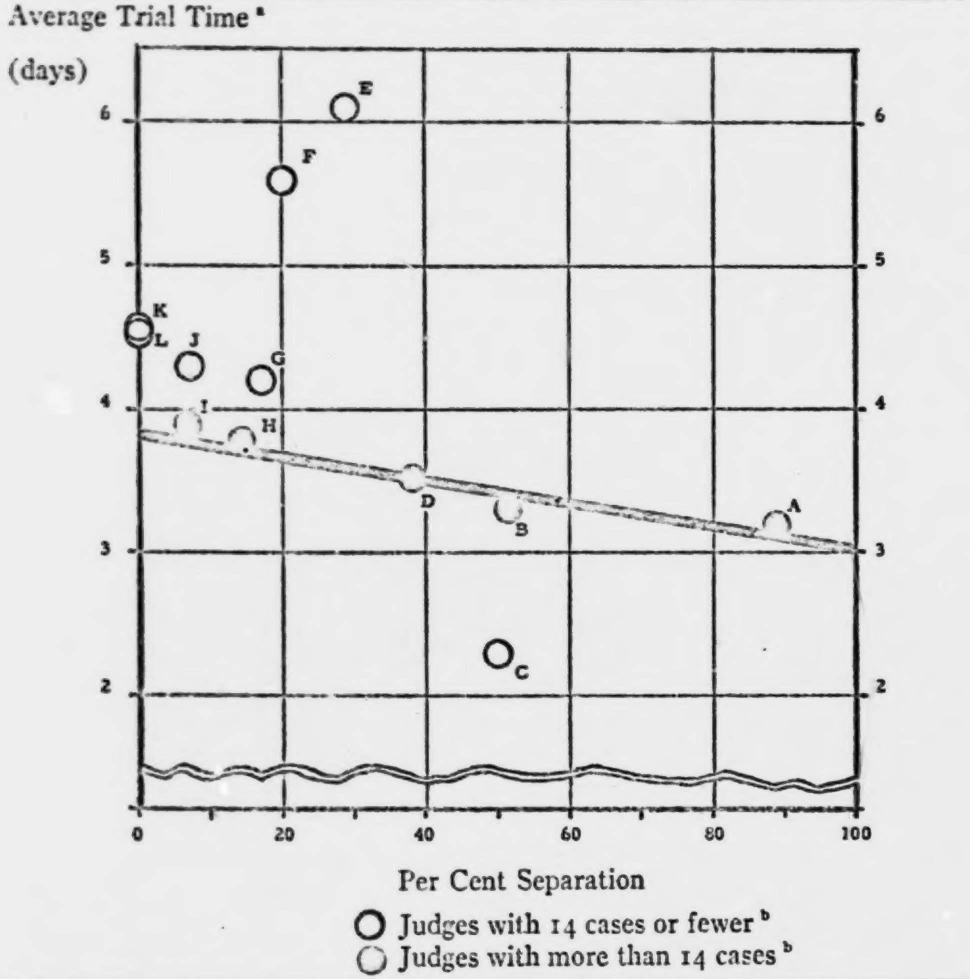
Judge	Proportion of Cases Tried Under Rule 21 (Per Cent)	Average Length of All Trials Before This Judge (Days)	Number of Trials Before This Judge
A[a]	89	3.2	(26)
B	51	3.3	(41)
C	50	2.3	(10)
D	38	3.5	(26)
E	29	6.1	(7)
F	20	5.6	(5)
G	17	4.2	(6)
H	14	3.8	(22)
I	7	3.9	(27)
J	7	4.3	(14)
K	0	4.5	(1)
L	0	4.5	(1)

[a] It might be relevant to reveal that Judge *A* was not Judge Miner, the proponent of the rule.

the shorter the average length of all his trials. In our case, we should expect Judges *A* and *B*, who have the highest separation rates, to show lower average trial lengths for all their trials than Judges *I*, *J*, *K*, and *L*. Table 4 presents the pertinent statistics. It is as we expected: as the separation ratio increases, the average trial length decreases. The exact nature of this relationship becomes even clearer if we put Table 4 into graph form as in Table 5.

TABLE 5

PER CENT SEPARATION AND TRIAL TIME



---

[a] In all personal injury jury trials, whether separated or not.

[b] Judges are identified by letter as they appear in Table 4.

Generally, Table 5 shows a band of dots sloping downward toward the right-hand borderline and thus confirms our expectation that the average trial length decreases as the proportion of separate trials per judge increases. As will be noted, the individual points representing the various judges are far from forming one line which would represent the relationship between separation and trial length; they show great individual differences. These differences derive from two causes. First, from the fact that some judges will try their cases more expeditiously than others; if so, their "points" will lie somewhat lower than they would if they tried their cases at average speed. Second, from the fact that, although cases are assigned to the judges at random, some judges will be assigned more time-consuming cases than others. The distorting effects of both these factors are automatically reduced to the extent that the number of judges (points) grows, so that the "slower" and "faster" ones balance each other — provided that there is no consistent correlation between "fast" judges and their propensity to separate.[12] The distorting factors are also reduced to the extent to which the individual judges have more cases and hence are likely to have on balance the same average assortment of long and short cases. In line with this reasoning, we will disregard all judges who had less than fifteen cases. If we then look at the five remaining Judges, *I*, *H*, *D*, *B*, and *A*, we find them to fall nicely about the straight line we have drawn in Table 5.[13] The two crucial points along this line are its beginning and its end. It begins (at separation zero) with 3.8 days and ends at the right-hand margin (separation 100 per cent) at 3.0 days. This line, the much-looked-for unbiased estimate, suggests that if the court had tried all its personal injury jury cases under the separation rule, it would have saved 0.8 days out of 3.8 days, or 21 per cent of the total trial time.

IV

An additional table provides further insights into the means by which time is saved through the separation process: Table 6 compares the stage and mode of termination of regular and separated trials. The figures from Table 2, telling us at what stage the cases terminate, form the right-hand margin column of Table 6; those from Table 3, mode of disposition, form the last lines for regular and separated trials.

---

[12] There is, in fact, no assurance that such a correlation does not exist. A sufficient number of nonjury trials for each judge, or adequate statistics on his trial time prior to rule 21, might provide a clue. In this important respect, then, this analysis deviates from a controlled experiment with its concomitant reassurance that no systematic bias interfered with the result.

[13] If the line were based on all 12 judges, irrespective of their number of trials, it would show a somewhat steeper slope; that line would signify greater potential savings than the present one. The context of this analysis, however, gives us good reason to believe that the present line is likely to be closer to the "true line" that would emerge from a controlled experiment.

TABLE 6

STAGE AND MODE OF DISPOSITION OF PERSONAL INJURY JURY TRIALS

Stage of Termination	Settled	Verdict for Defendant		Jury Verdict for Plaintiff on Liability and Damages	TOTAL
		Directed Verdict	Jury Verdict		
		A. Regular Trials			
During Plaintiff's Case	18%	..	..	..	18%
At End of Plaintiff's Case	3%	1%	..	..	4%
During Defendant's Case	..	..	..	..	..
After Full Trial	3%	2%	31%	42%	78%
Total	24%	3%	31%	42%	100%[a]
		B. Separated Trials			
During Liability Trial	10%[c]	9%	..	..	19%
At End of Liability Trial	15%[d]	4%	43%	..	62%
During Damage Trial	4%[d]	..	..	..	4%
After Full Damage Trial	3%[d]	..	..	12%	15%
Total	32%	13%	43%	12%	100%[b]

[a] 117 trials.

[b] 69 trials.

[c] Settled before jury returned any verdict.

[d] Settled after jury verdict affirming liability but before jury returned verdict on damages. These cases, totaling 22 per cent (15 + 4 + 3), represent those in which the jury found for the plaintiff on liability but never deliberated on the question of damages.

If we look first at regular trials, we see that 18 per cent of all regular trials were terminated through settlement during the plaintiff's case, and that settlement at a later stage is much less frequent. Then we turn to the figures for separated trials. Here we see confirmed our hypothesis that the bulk of cases disposed of at the end of the liability trial terminate in a verdict denying liability: of the 62 per cent terminating at that point, 47 or roughly three-fourths, end with verdicts for the defendant. But the unexpected element concerns those cases in which the jury finds for the plaintiff on the issue of liability. We find an increased willingness on the part of defendants to settle once they have lost the liability issue. Fifteen per cent of all trials end through settlement after the jury affirms liability, another 4 per cent are settled during the damage trial, and 3 per cent at the end of the damage trial. Thus, of the (15 + 4 + 3 + 12 =) 34 per cent of all cases where the jury finds for the plaintiff, only 19 per cent reach a trial on the damage issue and only 12 per cent, or about one-third, reach a second jury verdict. In short, separation may save trial time, not only in cases ending in defendants' verdicts but in cases resulting in plaintiffs' verdicts as well. Once liability is affirmed these cases too are very likely to be settled and, hence, will benefit from separation of the issues.

These figures explain why it would be so difficult to predict at the time of separation which cases are likely to benefit from it. A prediction that the case is likely to end for the defendant is not enough, because the cases that go for the plaintiff on liability also are likely to benefit from separation. To select effectively, the judge would have to be able to identify in advance the two groups of cases that will *not* benefit from separation. These are (1) the 15 per cent of cases which go through a full trial despite separation, and (2) those cases whose early termination, though after the liability trial, cannot be credited to separation, since they would have ended prematurely through settlement or directed verdict even if they had been tried regularly. The first two figures in the right-hand column of Table 6, adding up to (18 + 4 =) 22 per cent, are a measure of the frequency of these cases. We now appreciate why Tables 2 and 5 come out so similarly. The "purposeful" selection by the judge is, at least with respect to the expected time saving, not too different from a selection by lottery.

V

Now we may consider a third approach to the problem of estimating how much time is saved by the separation procedure. It is clear that the entire time saving comes from avoiding the litigation of damages, either in cases where liability is rejected, or in cases which are settled after liability is affirmed. We can then roughly estimate the time saved if we know three items: The proportion of trials that go their full course to verdict if separation is not available, the proportion of separated trials that go the full course to a second verdict, and the share which litigation of damages constitutes of the total litigation time in a full trial. We know the first percentage to be 62 per cent; [14] from Table 6 we learned that only 12 per cent[15] of the separated trials went to a second verdict. This means that separation causes about (62 − 12 =) 50 per cent of the trials to terminate without running their full course. As to the share of total litigation time devoted to the damage issue, we must rely on expert estimates, although theoretically one could produce an actual time count on a representative sample of trials. Conversations with judges and lawyers have led us to conclude that on the average the damage issue constitutes at the most 40 per cent of the total trial time. Armed with these figures, we can perform the following computation: If each of the roughly 50 per cent of the cases in which litigation of damages is avoided involves a saving of 40 per cent of trial time, then the saving from separation would amount to 20 per cent of the total trial time. This result, in spite of the roughness of its basis, is close enough to our other findings to permit now the firm generalization that in personal injury jury trials *separation saves trial time of the magnitude of about 20 per cent.*

---

[14] According to data furnished us by the Administrative Office of the United States Courts, in the two-year period prior to the introduction of civil rule 21, 62% of personal injury jury trials went to a jury verdict.

VI

An important limitation should now be noted on the amount of time which courts may actually save through the separation device. All calculations so far are based on the assumption that *all* cases would be tried under the separation rule. But, as Table 1 indicated, in this court only 37 per cent of the personal injury trials are separated, and hence the time saved thereby is not 20 per cent but only 37 per cent of this 20 per cent, that is (.37 × 20 =) 7.4 per cent of the total trial time. We also know, however, from Table 4, that this average of 37 per cent is not the same for all judges. We know from conferences with Judge *A*, who had 89 per cent of his trials separated, that he ordered separation as a matter of routine, unless cause was shown as to why separation should not be ordered. Judges *I* and *J*, on the other hand, apparently proceeded with regular trials unless a special reason was advanced for ordering separation. We would conclude, therefore, that 10 per cent is a good enough estimate of the proportion of cases which, under the most liberal application of the separation rule, would be left to the regular mode of trial. And, consequently, we conclude that if the separation rule were applied liberally the overall time saving would not reach 20 per cent but only (.90 × 20 =) 18 per cent.

---

[15] In addition, 3% of the separated cases were settled after a full trial. We do not include these cases in the figure for trials which had run their full course, however, since the figure for regular trials does not include cases which ended in a settlement or directed verdict after the defendant's case, and we wish to have the figures as comparable as possible. There remains of course the difficulty that the actually separated trials do not behave necessarily like trials selected at random for separation. But this error is likely to be small enough to be ignored here.

VII

With this we reach the end of our efforts to measure the amount of trial time saved through separation. The remainder of this essay is concerned with factors that could conceivably operate in the opposite direction, increasing the court's burden of trial time.[16] The most obvious one might be that the amount of time needed for the jury's deliberation in separated cases — where the jury may deliberate twice — may be greater than in regular trials. Second, separation could conceivably reduce the number of jury waivers, increasing the proportion of the more time-consuming trials before juries.[17] Third, separation might increase the number of hung juries, and hence the number of cases that have to be tried twice.[18] Fourth, separation might reduce the proportion of cases settled prior to trial,[19] and thereby increase the number of cases that need to be tried. The suspicion that separation of issues may lead to an overall slackening of trial speed has already been laid to rest, especially by Table 2, which showed that the separated trials that go to two verdicts do not last longer than full regular trials. We turn now to the investigation of the possible offsets enumerated above.

---

[16] Before proceeding, it is necessary to report on two minor housekeeping details. One concerns the court's use of impartial medical experts; the other, our treatment of 13 cases in which only the damage issue was tried because liability was stipulated.

At the time the separation rule was introduced the court put into effect another innovation. The judges in their discretion were enabled to appoint an impartial medical expert if the adversary experts were too far apart. Since it could be argued that such experts could affect trial time and, hence, our analysis, we give herewith the relevant data. They carry the strong suggestion that these experts, whatever their effect, did not distort our analysis of trial separation. Fourteen per cent of the separated trials, and 8 per cent of the regular trials, had such experts. The average length of these trials was 2.5 days for the separated and 5.3 for the regular trials. There is thus no significant deviation from the general pattern. In any event, as we have pointed out elsewhere, the primary effect of the impartial expert is likely to be his increasing the likelihood of settlement *before* trial. DELAY 122–24.

Concerning the 13 cases with stipulated liability, the question had to be answered as to whether these stipulations would have taken place if the separation rule had not existed. Defense counsel might have simply preferred not to let the jury know just how negligent his client had been. For purposes of our statistics we treated the cases as follows: If the judge separated the issues, we assumed that admission of liability was in response to the separation order. Out of the 13 cases where liability was admitted, 3 fell into this category. The remaining 10 cases were counted among the regular trials, under the assumption that here liability would have been admitted even if the case had reached trial prior to the adoption of the separation rule. This treatment probably concedes less influence to the separation rule than it deserves. The 10 cases, treated as regular trials, represent (10 out of 117 =) 9% of all regular trials. Although we have no statistics on such cases for the time prior to the rule's adoption, it seems unlikely — and the judges have confirmed this in their conferences with us — that this percentage was as high as that. We probably err, therefore, by crediting only 3 cases of admitted liability to the separation rule. This error will tend to make the average trial time for regular cases shorter, and for the separated case longer, than it actually was; hence, it probably causes us to underestimate the savings attributable to separation.

A. Increased Time for Jury Deliberation?

The first point is, in any event, one of relatively minor importance, since jury deliberation time is not necessarily court time. Nevertheless, it will be of interest to see the pertinent figures in the following table.

TABLE 7

AVERAGE JURY DELIBERATION TIME AND FREQUENCY

	Deliberation Time	×	Per Cent of Trials in Which Jury Deliberated	=	Weighted Average
Separated Trials					
Liability Issue	2.7 hours[a]		71%		1.92 hours
Damage Issue	3.4 hours		12%		0.41 hours
Total					2.33 hours
Regular Trials					
Both Issues	3.7 hours		73%		2.70 hours

[a] This is the average figure for all trials. In trials where the jury actually deliberated twice, it took only 2.2 hours for the jury to affirm liability.

Thus we see that the jury in separated trials, if it deliberates twice, spends an average of (3.4 + 2.2 =) 5.6 hours in the jury room as against an average of only 3.7 hours in regular trials. This could be an offset, except that in only 12 per cent of the separated trials the jury was called upon to deliberate twice.[20] The last column of the table takes into account the frequency of deliberation, giving an average deliberation time for all cases, including those in which the jury deliberated only once or not at all. With deliberation time thus weighted, the effect of separation is favorable rather than unfavorable; the table reveals that the jury spends on the average 0.37 hours, or about 22 minutes, less deliberation time in the average separated trial.

---

[17] See Weinstein, *Routine Bifurcation of Jury Negligence Trials: An Example of the Questionable Use of Rule Making Power*, 14 VAND. L. REV. 831, 851 (1961).

[18] See 74 HARV. L. REV. 781, 782 (1961).

[19] See Weinstein, *supra* note 17, at 850.

[20] See Table 6, p. 1617 *supra*.

B. Decreasing Frequency of Jury Waivers?

On the second point, the frequency of jury waivers, the evidence is equally reassuring. There is no ground for believing that the proportion of demands for juries increases under the separation rule; in fact the evidence seems to point the other way, though the difference is negligible. During the two-year period before separation, juries were used in 90 per cent of all personal injury cases;[21] during the corresponding period after separation, in only 86 per cent.[22] Furthermore, if separation were to reduce jury waivers, one would expect the judges who are more likely to order separation to show a lower incidence of jury waivers. But Table 8, which presents the relevant data, shows no evidence to this effect.

TABLE 8
SEPARATION RATIO AND JURY WAIVER IN PERSONAL INJURY TRIALS

Judge	Per Cent of Cases Tried Under Rule 21 [a]	Per Cent Waiver [a]	Number of Trials Before This Judge [a]
A	82	21	(33)
B	53	9	(45)
C	45	9	(11)
D	36	16	(31)
E	29	0	(7)
F	17	17	(6)
G	14	14	(7)
H	11	18	(27)
I	6	18	(33)
J	7	0	(14)
K	0	0	(1)
L	0	0	(1)

[a] These separation ratios differ from those in Table 4 because they are based on all trials conducted by this judge, not only on his jury trials.

C. Increase in Hung Juries?

As to the third possible offset against the savings achieved in trial time, a possible increase in the number of hung juries, they are as rare an event now as they were prior to the new rule.[23] But even without such specific evidence it would seem unlikely that any major offset could be expected from this source. Even if the ordinary 3 per cent rate of mistrials had been doubled under the new rule (which it was not), the offsetting loss of trial time would be only about 1.5 per cent, since a hung jury is a total loss only in about half of all instances; in the other half the litigants take the jury's hint and settle, so that there is no need for a retrial.

---

[21] Statistics supplied by the Administrative Office of the United States Courts.

[22] From Table 1, p. 1609 *supra*; 186 is 86% of 216.

D. Decrease in Settlement Ratio?

On the fourth issue, the settlement ratio, the evidence is given in Table 9, which shows that the proportion of cases reaching trial has not changed at all.

TABLE 9

SETTLEMENT RATIO IN JURY CASES BEFORE AND AFTER RULE 21

	Two-Year Period Before Separation		Two-Year Period With Separation	
	Number	Per Cent	Number	Per Cent
Disposed Before Trial	1218	84.9%	1217	84.9%
Reached Trial	217	15.1%	216	15.1%
Total Personal Injury Jury Cases Filed	1435	100.0%	1433	100.0%

That the separation rule is not likely to affect the settlement ratio can also be appreciated on a priori grounds. The likelihood that parties will settle is not a function of the relative strengths of their positions, but rather of the clarity with which each sees the other's position, since a settlement arises when each party can weigh the other's chances and agree with him on what those chances are. Whatever possible difference the rule could make to the verdict expectation of the litigants, there is no reason to believe that such a difference would be more visible to one side than to the other. We should expect, therefore, that any possible difference would leave the litigants as likely, or as unlikely, to settle as they had been before.[24]

---

[23] Unfortunately there are no precise court statistics available on the frequency of hung juries. But our inquiries with both judges and clerks yielded complete consensus on the insignificance of hung juries in this context.

[24] Compare the analysis of an analogous problem: *Interest From the Day of Accident*, DELAY 128-40.

VIII

Although our inquiry has taken us far afield and into several side forays, the task has been a narrow one — to measure the effect of separation of issues on the court's trial load. This then is a summary of our findings:

Separation of issues will save, on the average, about 20 per cent of the time that would be required if these cases were tried under traditional rules. This saving derives from the fact that in many cases separation makes the litigation of damages unnecessary. This group includes all cases in which liability is denied, but also the majority of cases in which liability is affirmed, because two out of three of these cases are likely to be settled without trial of the damage issue. There is no evidence that this saving is offset by a change in the settlement ratio prior to trial, in the frequency of jury waivers, or in the proportion of hung juries, any one of which factors — if affected — could increase the court's trial load.

It is not possible to sort out effectively in advance the cases in which separation would prove futile. Therefore, if a court wants to realize the maximum of potential time saving through separation, it should separate as frequently as possible. The time saved will be in direct proportion to the frequency of separations. Such a policy recommends itself also on the ground that separation is unlikely ever to *add* substantial trial time even if cases should go, as some do, to a second verdict. If a court were to follow Judge *A*'s example and order separation routinely unless cause is shown to the contrary, it could expect separation in about 90 per cent of all cases. On this level of separation the overall saving could be expected to amount to roughly 18 per cent of the court's trial time.

Viewed superficially, the route which took us to this result might seem overlong. Yet such is the nature of the social sciences that, short of a controlled experiment, all their evidence is imperfect; only through a variety of converging approaches can one hope to reach safe ground.[25] But from our investigation we have learned that separation is a powerful remedy for court congestion. Fully used, it would be equivalent to increasing the number of judges trying those cases by one-fifth. Cases which, if not separated, will require five judges for their disposition will, if separated, require only four judges and free one for other work.

The introduction of the separation rule, as we had occasion to observe, raised other questions beyond its efficacy as a delay remedy. These questions concern the possible effect of separation on the substance of the jury verdicts. By depriving the jury of its joint verdict, it is argued, subtle influences operate to affect the verdict, and should therefore be included in an overall appraisal of the rule. We are now in the process of collecting data which might have a bearing on this problem; if findings prove significant, they will appear in a subsequent report.

---

[25] *Cf.* Zeisel, *Social Research on the Law: The Ideal and the Practical*, in LAW AND SOCIOLOGY 124–43 (Evan ed. 1962), especially p. 126.

QUALITATIVE ANALYSIS

The largest share of our attention has been devoted to quantitative data rather than qualitative data. Our emphasis has been due more to the complexity of quantitative analysis than to judgments about the value of qualitative versus quantitative materials. In fact, we feel that qualitative material is valuable to the social sciences and that any discussion of social science methodology would be lacking if the student were not made aware of its value.

In general, qualitative research is used in two capacities: (1) it serves as a first step toward quantification by discovering "unexpected phenomenon which stimulate a search for causal explanations;" and (2) it offers the researcher a different kind of approach, a "substitute for statistical research making the same kind of statements but on the basis of recording and analysis . . . which takes place largely within the mind of the observer." The following paper by Allen Barton and Paul Lazarsfeld is an excellent statement of the role of qualitative analysis in social research and reveals its great importance. The great bulk of man's behavior, of course, is still far from being easily measurable--if, indeed, many of its most significant aspects can ever be measured in the traditional sense.

There is one particular aspect of the following selection which is particularly worth noting. Qualitative data, like quantitative data, can be systematically ordered and analysed even though it does not offer the same precision. As Barton and Lazarsfeld point out, the researcher must organize the raw observations into a descriptive system. These descriptive systems, as the authors indicate, can range from crude lists of types, each defined individually to fully systematic typologies in which each type is a logical compound of a small number of basic attributes. What all such typologies have in common, irrespective of their complexity, is that from them one can derive a classification of the values, habits, and attitudes which are important to the explanation of social behavior. Thus Louis Wirth's typology of the Jewish community, involving the social roles of Mensch, the Allrightnick, the Schlemihl, the Luflmensh, the Yeshiba Bochar, and the Zaddick, give us as complete an index as any obtainable of the culture traits and the culture pattern of the group.

From Allen H. Barton and Paul F. Lazarsfeld, "Some Functions of Qualitative Analysis in Social Research," Frankfurter Beitrage Zur Soziologie (Band 1, 1955), pp. 321-361. (Reprinted from Bobbs-Merrill Reprint Series in the Social Sciences, #S-336.) Reprinted by permission.

Allen H. Barton and Paul F. Lazarsfeld, "Some Functions of Qualitative Analysis in Social Research," Frankfurter Beitrage Zur Soziologie,(Band 1, 1955), pp 321 - 361. (Reprinted in the Bobbs-Merrill Reprint Series in Social Sciences, Number 5-336).

The advancement of research procedure in social science as elsewhere depends on making explicit what researchers actually do, and systematically analyzing it in the light of logic and of substantive knowledge. Such a "codification" of procedures points out dangers, indicates neglected possibilities, and suggests improvements. It makes possible the generalization of methodological knowledge – its transfer from one specific project or subject matter to others, from one researcher to the scientific community. Finally it makes possible a more systematic training of students, in place of simply exposing them to concrete cases of research in the hope that they will somehow absorb the right lessons.

Such a recording and analysis of procedures has gone quite far in certain parts of the social research process – in the design of experiments, in the analysis of survey data, in the scaling and measurement of social and psychological variables, and in sampling. But codification has been very unevenly applied; important parts of the research process have been neglected.

This is particularly true of the analysis of non-quantified data, "qualitative analysis", as it is often called. A great deal of social research operates with qualitative descriptions of particular institutions, situations or individuals, rather than with "largely quantified data accumulated by structured observation in empirical situations approximating (with specified deviations) the model of controlled experiment."[1] Not only is this type of research large in volume, but it plays important roles in the research process, by itself and in connection with quantitative research. This paper aims to make a start at the systematic analysis of "qualitative procedure".

The question which we would like to answer is: "What can a researcher do when confronted by a body of qualitative data – detailed, concrete, non-metric descriptions of people and events, drawn from direct observation, interviews, case-studies, historical writings, the writings of participants?" The methodologist's first step toward a systematic answer is to examine what researchers in

---

[1] Definition of quantitative research given by Daniel Lerner, "The American Soldier" and the Public, in: Robert K. Merton and Paul F. Lazarsfeld, Continuities in Social Research, Glencoe, Illinois, 1950, p. 220.

POOR ORIGINAL COPY - BEST
AVAILABLE AT TIME FILMED

fact have been doing with qualitative material. About 100 studies were culled for characteristic examples. An effort was then made to organize these cases in order that the most characteristic types of qualitative work could be distinguished and documented. This paper presents the resulting organization.

The reader will have no difficulty in noting that this "guide through qualitative research" is itself guided by proceeding from simple to ever more complex procedures. We begin with a discussion of the value of simple observations. We then proceed to those studies which center on ordering and classification. Our next group of examples demonstrates the various ways in which several variables are interrelated through qualitative analysis. Next we discuss cases where the analyst wants to encompass such a great number of dimensions that he cannot make them all explicit, but tries to sum them up in a general "pattern". This is probably the point at which qualitative research is most creative, most controversial, and most difficult to describe. It will be seen that we had to use a special term ("matrix formula"), to bring into relief this means of seeing the social world in a new way. Finally, we touch on the role of qualitative data in the support of theory, a topic so large that we did not dare to pursue it to any extent.

It should also be kept in mind what this paper does not attempt to do. First of all, it does not describe how qualitative research should be done; it is restricted to an organized description of what is actually being done, without expressing any judgment. Secondly, this paper, but for one point, does not make any attempt at formalization. The exception is in Section II where the use of typologies is discussed. The logic of typologies is by now so well developed that it was simple to include it in this paper. Such formalizations have considerable advantages. They indicate the underlying assumptions in a given piece of qualitative work, what points the author might have overlooked, at what points he might have contradicted himself, and so forth. There is no doubt that additional formalizations will be needed. We have, for example, distinguished various ways in which a single observation can be fruitful; however, we have not tried to bring the different possibilities into a more general context from which they can be derived. As a matter of fact, one of the hopes for the present survey is that it will facilitate further work in this direction[1].

---

[1] The general idea of codifying existing methods of social research is now being carried out by a special project on Advanced Training in Social Research at Columbia University. Formalizations of specific pieces of qualitative writing are an essential part of this project. Available in monograph form so far are a formalization by Ernest Nagel of procedures of functional analysis in social research, and a formalization by Paul F. Lazarsfeld of certain problems of process analysis, the interaction of social variables through time. The present can be considered as the typical first step which has to precede a more full-fledged formalization. It is one of the documents developed by the Columbia project.

We have also not dealt with the problem of evidence. Under what conditions in the social sciences an assertion is proved is a very difficult question, not restricted to qualitative research. It seemed best not to touch on this issue, in a context in which all emphasis was placed on providing a picture of a kind of work which is usually considered so "private" that it defies all systematic presentation.

I. ANALYSIS OF SINGLE OBSERVATIONS

When one examines qualitative reports, one of the first types of material which catches our attention is the "surprising observation". Like the nets of deep-sea explorers, qualitative studies may pull up unexpected and striking things for us to gaze on. We find that there are people who believe that they are being educated by the unrelated and trivial information presented by quiz shows[1]. Interviews with people deprived of their newspaper by a strike disclose that some do not turn to alternative sources of news, but to reading anything which is lying around the house; a major function of newspaper reading seems to be simply to fill in "gaps" in the daily routine[2]. Observers of the underworld tell us that professional thieves constitute a rather exclusive social group, with exacting standards of membership strongly reminiscent of those of lawful professions[3]. Anthropological data of course are full of surprising observations: that Eskimos lend their wives to guests without any jealousy, that Fiji Islanders kill their chiefs when they grow old, and so on[4].

These phenomena are of various levels: some are individual beliefs and behaviors, some are a matter of group standards and structures within a society, some involve the norms of a whole culture. In each case the qualitative researcher has simply disclosed that such-and-such a phenomenon exists. And in one way or another, to be told that such things exist has a strong impact on the reader. They all have an element of surprise.

In the next few pages we will try to clarify what this impact is, and what research functions are served by these qualitative observations which simply state the existence of something surprising. We can distinguish at least two different uses for such observations. First, the existence of a phenomenon may raise problems – that is, compel us to look for explanations, to explore its consequences, to try to fit it into our scheme of knowledge. Second, we may find in the qualitative observations an indicator of some general variable which we want to study, but cannot measure directly.

---

[1] Paul F. Lazarsfeld, Radio and the Printed Page, New York 1940, pp. 74 seq.

[2] Bernard Berelson, What "Missing the Newspaper" Means, in: Paul F. Lazarsfeld and Frank Stanton, Communications Research 1948–1949, New York 1949, pp. 122 seq.

[3] E. H. Sutherland, The Professional Thief, Chicago 1937.

[4] A variety of anthropological examples is given in Margaret Mead, Adolescence in Primitive and in Modern Society, in: Theodore M. Newcomb and Eugene L. Hartley, Readings in Social Psychology, New York 1947, pp. 6 seq.

Observations Which Raise Problems

Some observations are surprising because they conflict with our expectations, either common-sense or theoretically derived. Other observations surprise us by bringing to light phenomena which are simply new and unexplained, which challenge our curiosity. Yet another important type of problem-raising observations is that which brings together under a clear label a body of "familiar" experiences which had not previously been seen as a definite, generally occurring social phenomenon – which forms for the first time, so to speak, a "social object" to be studied.

In any of these cases the result is that a problem is raised. Our attention is focussed on a phenomenon, and we are stimulated to seek explanations and inquire into consequences. To make such a problematic observation is to initiate a research process which may lead to significant advancement of our understanding of social phenomena. (Some kind of observations no doubt raise more significant questions and lead to more valuable findings than others. It would be of great value to develop "screening principles" which can direct our attention to the more significant of the surprising observations; this cannot however be gone into here[1].)

To give concrete meaning to the notion of a problematic qualitative observation, and to provide material for its further development, a number of examples can be given.

As is well known, the original experiments of the Western Electric researchers led to highly surprising quantitative results: the experimental changes in physical conditions of work in no way accounted for the changes in production in the experimental group or workers. At this point the researchers decided to go back to the very first stage of the research process, and simply gather observations about what goes on in a normal working group in a factory. This exploratory research turned up a number of surprising qualitative observations[2]:

> Some work groups were characterized by "a lack of ambition and initiative and a complacent desire to let well enough alone";

---

[1] A discussion of this point is found in Robert K. Merton, Social Theory and Social Structure, Glencoe, Illinois, 1949; Chapter III, The Bearing of Empirical Research on Sociological Theory.

[2] F. J. Roethlisberger and William J. Dickson, Management and the Worker, Cambridge, Mass., 1946, Chapter XVII.

> "The supervisory control which is set up by management to regulate and govern the workers exercises little authority . . ."
>
> "They (the employees) firmly believe that they will not be satisfactorily remunerated for any additional work they produce over the bogey . . ."

All of these facts were in conflict with what the management and the researchers had expected. By following them up the researchers were led to their now well-known discovery of the importance of informal groups in formal organizations, and of the barriers to communication which exist between levels within organizations.

Communications research offers many examples of surprising qualitative observations. We have already mentioned the discovery of unexpected motives – for listening to quiz programs and for reading newspapers. Unexpected responses to communications are another important example. A broadcast warning the public to patronize X-ray operators and avoid "quacks" left some listeners afraid of any X-ray treatments, and others doubting whether there could be any X-ray machines in the hands of incompetent operators[1]. A film designed to impress Americans with the British war-effort left some more convinced than ever that America was bearing all the burden[2]. The discovery of such anomalous responses led to more detailed investigations of the communications process, which turned up some important general principles – for example, about the need to relate the message to the experience-world for the audience, which may be quite different from that of the communicator.

A study of how prejudiced people respond to cartoons ridiculing prejudice found an unexpected type of response[3]. Some people were neither shamed out of their prejudices nor insulted; they simply did not understand what the cartoons were driving at. When this response was investigated in detail, the "derailment mechanism" of motivated misunderstanding was revealed (see p. 338 below).

Listeners to Kate Smith's war-bond "marathon broadcast" placed remarkable emphasis on her "sincerity"[4]. Considering that other professional entertainers drew no such response, and that the respondents were generally suspicious of advertising and propaganda manipulation, this seemed worth investigating. Further study suggested the importance of "the propaganda of the deed" in a propaganda-wary society – in this case, Kate Smith's presumed strain and sacrifice in making the 18-hour broadcast.

---

[1] Paul F. Lazarsfeld and Patricia L. Kendall, The Listener Talks Back, in: Radio and Health Education, New York 1945.

[2] Robert K. Merton and Patricia L. Kendall, The Boomerang Response, in: Channels, Vol. XXI, No. 7 (June 1944).

[3] Patricia L. Kendall and Katherine M. Wolfe, The Analysis of Deviant Cases in Communications Research, in: Lazarsfeld and Stanton, l. c., pp. 158 seq. Also E. Cooper and M. Jahoda, The Evasion of Propaganda, Journal of Psychology, Vol. XXIII, 1947, pp. 15 seq.

[4] Robert K. Merton, Mass Persuasion, New York 1946, pp. 82 seq.

The examples so far have involved unexpected phenomena which stimulated a search for causal explanations. In other instances the problem which is raised is in the other direction – an investigation of the consequences of a certain phenomenon is stimulated. Thus one researcher interested in problems of the profession noted an "obvious" fact as raising a problem: while all professionals meet a certain proportion of failures, the trial lawyers as a group necessarily lose half their cases[1]. What must be the consequences of such a high rate of failure for these professionals, and how do they deal with it? The answers to these problems might throw light on some important problems of the professional role. The ability to take a commonplace fact and see it as raising problems is important because it can lead ultimately to such enlightenment.

Another such observation was made by Merton in his study of the Kate Smith war-bond marathon. In the content of the broadcast there was no reference to the real economic purpose of war bond buying as an anti-inflation measure. Merton saw this fact as raising problems of consequences: an opportunity to educate a large and attentive audience in economic realities had been neglected, and reliance had been placed instead upon "large delusive statements" playing upon the anxieties of those with loved ones overseas. What were "the further, more remote but not necessarily less significant effects of these techniques upon the individual personality and the society?" "Does the unelaborated appeal to sentiment which displaces the information pertinent to assessing this sentiment blunt the critical capacities of the listeners?"[2]

An example of the forming of commonplace experiences into a clearly labelled social phenomenon, and thereby creating a new object for investigation, is Adler's formulation of the concept of the inferiority complex[3]. Everyone at one time or another has experienced feelings of being inadequate, unworthy, etc., but until these private sensations had been pointed out and labelled, they could not be investigated by social science. Other examples which might be called to mind are Veblen's formulation of the concept of conspicuous consumption[4], or Sutherland's labelling of certain categories of business behavior as "white collar crime"[5]. Without having actually discovered any new facts, simply by directing attention to familiar facts placed for the first time within a distinctive category, these investigators were able to raise important problems and initiate fruitful study.

---

[1] Wagner Thielens, research in progress, Columbia University, Department of Sociology.

[2] Merton, Mass Persuasion, l. c., pp. 188 seq.

[3] Alfred Adler, The Neurotic Constitution, New York 1926.

[4] Thorstein Veblen, The Theory of the Leisure Class, Modern Library, New York 1934.

[5] Edwin H. Sutherland, White-Collar Criminality, in: American Sociological Review, Vol. V, February 1940, No. 1, pp. 1 seq.; – White Collar Crime, New York 1949.

The reader may have noticed that some of the examples of "surprising observations" cited are no surprise at all to him. This is as it should be. The point is that one time they were surprising, and initiated further investigation which has been sufficiently successful to render them familiar and understandable today. Another problem is that an observation may be surprising to the particular researchers, while other social scientists have known about it all along. Thus the Western Electric researchers made the "surprising discovery" that informal social organization existed among workers. Other sociologists had long been aware of the problem of informal groups; however, the special preconceptions of American industrial sociology up to that time had kept it unaware of this whole realm of phenomena. In a still more extreme case, a finding may be new only to one particular researcher; in which case it might better be termed "self-education" than a scientific discovery.

Is there anything which a researcher can do toward making "surprising observations" other than to maintain an alert state of mind? It may seem contradictory to speak of giving instructions for making surprising observations. However, there are ways in which one can at least increase the probability of making such observations. Some of these are discussed by Jahoda, Deutsch, and Cook under the heading, "The Analysis of 'Insight-Stimulating Cases'."[1] Strangers or newcomers to a community or a country, it is suggested, may be able to pick out problematic facts which are simply taken for granted by those accustomed to the locale. Marginal individuals, or cases in transition from one stage or status to another, may present much more clearly certain problematic features of a personality-type of social system. Deviants, extreme cases, or "pure, ideal-typical" cases may have a relatively high efficiency in indicating problematic facts.

Observations Which Serve Indicators

The first type of surprising observations discussed were those which were anomalous and unexplained, which served the function of stimulating a search for explanations. Another type of qualitative observations are challenging because we see in them indications of some large-scale phenomenon which we cannot perceive directly. Thus the occurrence of riots and protest meetings in the North during the Civil War serves as an indication that opposition to Lincoln's war policy existed; bits of shell or pottery found in graves mark the routes of trade and cultural contact in the prehistoric world; a peculiar military custom indicates the caste-like nature of army organization; modes of speech may indicate complex mental patterns or cultural emphases.

---

[1] Marie Jahoda, Morton Deutsch and Stuart W. Cook, Research Methods in Social Relations. New York 1951, pp. 42 seq.

Three situations can be pointed out in which one pays particular attention to qualitative indicators. They are distinguished in terms of the kind of obstacle which prevents direct observation and measurement of the underlying variable:

(1) Situations in which qualitative evidence substitutes for otherwise simple statistical information relating to past ages or inaccessible countries.
(2) Situations in which qualitative evidence is used to get at psychological data which are repressed or not easily articulated – attitudes, motives, assumptions, frames of reference, etc.
(3) Situations in which simple qualitative observations are used as indicators of the functioning of complex social structures and organizations, which are difficult to subject to direct observation.

The underlying assumption in all these cases is that a phenomenon which cannot be directly observed will nevertheless leave traces which, properly interpreted, permit the phenomenon to be identified and studied. A great historical movement, a basic personality characteristic, an essential characteristic of organizational structure, should all leave their imprint on almost any documentary material, accounts by observers, or even physical refuse, which they leave behind.

Examples of the first class (qualitative substitutes for unavailable statistical or descriptive material) would include the use of newspaper stories or other contemporary records of public demonstrations as indications of public opinion in past times of crisis[1]; Frazer's use of advertisements for slaves in ante-bellum Southern newspapers to find out about the structure of the slave family[2]; the use of reports of refugees, Soviet press materials, and the contents of Soviet literature to provide information about life in the Soviet Union[3]; the use of archeological remains to indicate culture contacts or religious beliefs in prehistoric times.

Examples of the second class (qualitative indicators of psychological variables) include formal projective testing, the psychological analysis of personal documents or artistic works, the analysis of items of literature or entertainment as presumed projections of the traits of their audience[1], and of course the analysis of qualitative interviews or records of participant observation. A good example of the use of indicators for a psychological concept is found in the study of anti-prejudice cartoons mentioned earlier:

---

[1] For a number of examples, see Paul F. Lazarsfeld, The Obligations of the 1950 Polister to the 1984 Historian, in: Public Opinion Quarterly, Vol. 14, No. 4 (Winter 1950–51).

[2] E. F. Frasier, The Negro Family in the United States, Chicago 1939, pp. 55 seq. This and other examples are discussed in: Robert Bower, Training Guide on the Qualitative Use of Documentary Material, New York, Bureau of Applied Social Research, 1950; mimeographed.

[3] For example: Barrington Moore, Jr., Soviet Politics: the Dilemma of Power, Cambridge, Mass., 1950; Alexander Gerschenkron, A Neglected Source of Economic Information on Soviet Russia, in: American Slavic and East European Review, Vol. IX (February 1950).

> "In tracing the process through which these 68 respondents arrived at their misunderstanding, we find our starting point in the fact that most of them *identified* with Mr. Bigott. (Footnote:) By 'identification' we mean the mental process through which a subject assumes the role of another person to such an extent that actions, either verbal or behavioral, directed toward the object of identification are experienced as directed toward the identifying person. Evidence of identification with Mr. Bigott was manifested by the subject's acting in one or more of the following ways: (a) explicitly affirming identification, saying, for example, 'I guess I'm a Mr. Bigott'; (b) consistently and openly sympathizing with Mr. Bigott, expressing sorrow, for example, that Mr. Bigott looked so weak and sick in the 'transfusion' cartoon; (c) interpreting a threat to or criticism of Mr. Bigott as referring to himself, as, for example, becoming emotionally upset by the cobweb on Mr. Bigott's head."[2]

A single conversation reported by the authors of *Deep South* bears witness to the depth of feeling involved in white attitudes toward Negroes in this deeply prejudiced area: A social worker described a poor-white family in which two girls

> "'are having babies and are not married . . . That isn't the end by any means . . . Somebody told me that this older girl was sleeping with the father . . .' After the recitation of the case, when the social worker was out of the room, a woman whispered to the interviewer: 'Mrs. Wilson says *those girls have Negro men too,* but Miss Trent (the supervisor) won't let me say anything about that . . . *Isn't that awful?*' It is significant to note from this interview that the incest situation was viewed with les horror than the infraction of the caste sex taboos."[3]

On the basis of his long participant observation, William Whyte was able to report the following striking indicator of the complete acceptance of gambling in Cornerville:

> "When a mother sends her small child down to the corner for a bottle of milk, she tells her to put the change on a number."[4]

Investigators studying the effects of unemployment on the psychology of the people of an Austrian village had the school children write essays on the theme,

---

[1] For one example: Donald V. McGranahan and Ivor Wayne, German and American Traits Reflected in Popular Drama, in: Human Relations, Vol. I, No. 4 (August 1948), pp. 429 seq.

[2] Kendall-Wolfe, l. c., p. 163.

[3] Allison Davis, Buleigh B. Gardner and Mary R. Gardner, Deep South, Chicago 1943, p. 116.

[4] William F. Whyte, Street Corner Society, Chicago 1943, p. 116.

"My Future Occupation". The pervasiveness of the insecurity of the children of the unemployed, its corrosive effect on planning for the future, was indicated among other things by the very language used. Children of employed workers would write, "I will be ..." or "I want to become ..." Children of the unemployed tended to use phrases like "I might become" or "I would like to be ..." In the same study a small boy remarked to one of the investigators that he would like to be an Indian chief, "But I am afraid it will be hard to get the job."[1]

The third situation – the use of simple qualitative indicators to show the attributes of complex social structures – is very clearly exemplified in Blumenthal's study of a small mining community. The speed and inclusiveness of interpersonal communication in the community was indicated

> "by the fact that should a death occur at nine o'clock in the morning and the information not reach a resident until late in the afternoon, his usual expression is, 'I can't understand why I didn't hear that sooner', and others say to him, 'Where have you been? Everybody knew that by noon."[2]

At another point Blumenthal notes the existence of conflicting qualitative indicators of the social contact between Mineville and its nearest neighbor, and concludes that one has the greater weight:

> "During the heyday of Crystal her people and those of Mineville were not so well acquainted as might be supposed ... Hotly contested baseball games and the communities having celebrated together on the main day of festivities for each – Miner's Union Day – were not indications of far-reaching personal relations. This is shown by the measure of social distance evidenced by the fact that a young man whose reputation was such in one town that its 'respectable' girls refused to associate with him could go to the other and fraternize with its 'best' young woman."[3]

The existence of primary group relations within smaller units of the American army can be inferred from the following qualitative indications drawn from an interview:

> "We bunked together, slept together, fought together, told each other where our money was pinned in our shirts ... If one man gets a letter from home the whole company reads it."[4]

The authors of *The American Soldier*, wanting an indicator for the complex notion of the "Army caste system", pointed to an institutionalized symbolic act:

---

[1] M. Jahoda and H. Zeisel, Die Arbeitslosen von Marienthal, Leipzig 1932, quoted in: Jahoda, Deutsch and Cook, l. c., pp. 293 seq.

[2] Albert Blumenthal, Small Town Stuff, Chicago 1932, pp. 136 seq.

[3] l. c., p. 30.

[4] Samuel Stouffer et al, The American Soldier, Princeton 1949, Vol. II, p. 99

> "Enlisted men selected for officer candidate school were first discharged from the Army and then readmitted in their new and very different status."

Just as it is impossible to move from one caste to another in an ethnic case situation, so an enlisted man about to become an officer must leave the Army system before reentering in his new status. The continuation of this custom is a certain indication of the continuation of the attitudes of a "caste system" in the Army[1].

In discussing the family structure found in Middletown's various classes, the Lynds suggest a possible indicator of the position of the husband in the familiy:

> "It may not be wholly fantastic to surmise that there may be some significance for the understanding of local marital association in the hierarchy of terms by which local women speak of their husbands. There is a definite ascent of man in his conjugal relations as one goes up in the social scale, from 'my old man' through 'the man', 'he' (most frequent), 'the mister', 'John', 'my husband', to 'Mr. Jones'. The first four are the common terms among the working class families, and the last two among business class families."[2]

The indicators which have been referred to are of many different forms. Some are linguistic, some are symbolic acts, some are documentary, some are physical objects. As substantive knowledge of linguistics, social organization, and technology are applied to the problem, one may expect ever more sensitive and reliable interpretations of such qualitative indicators. To what extent interpretation of indicators will have to remain an art, and to what extent it can be made a science, is one of the important problems of qualitative research which we cannot attempt to discuss here.

II. CONSTRUCTION OF DESCRIPTIVE SYSTEMS

The previous section discussed what can be done with a single "point" of qualitative data; the present section considers what one does when confronted by a whole array of qualitative observations. As a first step toward understanding a field of human activity, one must organize the raw observations into a descriptive system. In some cases one has only to apply categories already set up by previous investigators or by the society itself, and proceed with the further stages of analysis. In other cases previously existing categories are clarified and revised by the attempt to apply them to a concrete body of data.

---

[1] l. c., Vol. I, p. 56, fn. 2. This example and others are discussed by Patricia L. Kendall and Paul F. Lazarsfeld, Problems of Survey Analysis, in: Merton and Lazarsfeld, l. c., pp. 183 seq.

[2] Robert S. and Helen Merrell Lynd, Middletown, New York 1929, p. 118, fn. 2.

And in some cases the researcher must create his own classification system for the material under study. It is this latter case which will be particularly considered here.

In terms of their formal structure, the descriptive systems created by investigators can range from crude lists of "types", each defined individually without clear logical relationship to the others, to fully systematic typologies in which each type is a logical compound of a small number of basic attributes. Between these end points are all intermediate degrees of "partial" systematization, including some sets of types which include in their definition virtually all the logical elements necessary to set up a multi-dimensional "attribute space", but in which the logical analysis has not been explicitly made. Descriptive systems may also vary in terms of their degree of concreteness or generality. A fully systematic typology may be based on dimensions of a highly limited, concrete nature, while a preliminary classification can be broad and general.

Preliminary Classifications

A classification which falls toward the unsystematized end of the continuum can be called a preliminary one, since it represents an essential first step toward the ideal of a fully systematic one. The importance of this first step from completely unordered data to a preliminary classification must never be underestimated. Until the data are ordered in some way, the analysis of relationships cannot begin; more refined categories normally develop out of the attempt to analyze relationships between preliminary categories; there is an interacting process between refinement of classification and the analysis of relationships.

A good preliminary classification must provide a workable summary of the wealth of elements in the original data, and include – even if in unsystematic form – the basic elements necessary for understanding the situation. A bad preliminary classification is one which is poor in elements and suggestiveness, which omits so many important aspects of the situation that analysis reaches a dead end, and one must go back to the original data for a new start. So long as the essential elements are suggested somewhere in the initial classification, they can be picked and recombined more logically as the analysis proceeds. The question of what it takes to make fruitful preliminary categories – whether the process can be systematized and taught, or whether it is wholly an individual art – is one of those which most needs exploring. The present discussion can only raise this question, and present a number of examples of the process.

Good examples of the use of this form of qualitative analysis – the formation of relatively unsystematized but fruitful classifications of people and situations – can be found in the work of the Chicago urban sociologists. Such an instance is Louis Wirth's suggestive notes on "some Jewish Types of Personality". Wirth defines his purpose in using this technique as follows:

> "The sociologist, in transforming the unique or individual experience into a representative or typical one, arrives at the social type, which consists of a set of attitudes on the part of the person toward himself and the group and a corresponding set of attitudes of the group toward him . . . The range of personality types in a given social group is indicative of the culture of that group."[1]

Wirth's gallery of "characteristic and picturesque personalities that are met with in the average community" includes:

> the *Mensch*, a person of superior economic status who has "achieved his success without sacrificing his identity as a Jew";
> the *al[illegible]gotnick*, who "in his opportunism, has thrown overboard most of the cultural baggage of his group";
> the *Schlemihl*, who belies the stereotype of the Jew as "the personification of the commercial spirit" by being "quite shiftless and helpless, failing miserably in everything he undertakes";
> the *Luftmensch*, who moves easily from one unsuccessful project to another, and whose "only apparent means of subsistence is the air he breathes";
> the *Yeshiba Bochar*, literally the talmudical student, the young man whose learning gives prestige irrespective of wealth or origin;
> the *Zaddik*, the "pious, patriarchal person . . . whose exemplary conduct is pointed to as an example"; and so on[2].

The purpose of presenting these types, drawn largely from the folklore and literature of the subject group, lies in the fact that

> "they are as complete an index as any at present obtainable of the culture traits . . . and the culture pattern of the group . . . Together they constitute the personal nuclei around which the fabric of the culture of the group is woven. A detailed analysis of the crucial personality types in any given area or cultural group shows that they depend upon a set of habits and attitudes in the group for their existence and are the direct expressions of the values of the group."[3]

Starting from these types, therefore, on can derive a classification of the values, habits and attitudes which are important to the explanation of the behavior of the group.

In much the same way, from folklore and literature as well as personal observation, C. Wright Mills draws a gallery of "white-collar types." There are types of managers:

> the "glum men" on the top of the white-collar pyramid, harrassed, cautious, careful to stay in line with the aims of the employers or other higher ups;

---

[1] Louis Wirth, Some Jewish Types of Personality, in: Ernest W. Burgess, The Urban Community, Chicago 1926, p. 106.

[2] l. c., paraphrased, pp. 108 seq.

[3] Ibid.

> the "old veterans" just below the top, who seek security in closely following explicit instructions, and strive for deference from those below;
> the "live wire", the younger man on his way up;
> the "new entrepreneur", who prospers as a fixer and go-between in a world of huge and complicated organizations, mass manipulation, and general insecurity[1].

There are types of intellectuals[2], of academic men[3], and – all the way down at the bottom of the white-collar pyramid – of salesgirls: "the wolf", "the charmer", "the ingenue", "the social pretender", and so on[4].

There is a serious purpose in pin-pointing these picturesque types:

> "By examining white-collar life, it is possible to learn something about what is becoming more typically 'American' than the frontier character probably ever was. What must be grasped is the picture of society as a great salesroom, an enormous file, an incorporated brain, a new universe of management and manipulation. By understanding these diverse whitecollar worlds, one can also understand better the shape and meaning of modern society as a whole, as well as the simple hopes and complex anxieties that grip all the people who are sweating it out in the middle of the twentieth century."[5]

The general run of preliminary categories will not be as colorful and rich in suggestions as these just quoted, but they will be of the same formal nature: a simple list of discrete "types". Thus we will have lists of "types" of comic-book readers[6], types of client-professional relations, types of appeals in a certain propaganda broadcast[7], types of communities, etc., representing a preliminary ordering of material into a simple list of headings. As the analysis progresses, either within the original study or in the work of replication or secondary analysis, these simple lists may be developed into more systematic and more general descriptive systems.

Somewhat further along the road to generality and systemization are the kinds of "types" found in the great deal of the speculative and theoretical literature. Typical examples here are Spranger's six "value types" – the theoretical, economic, aesthetic, social, political, and religious[8] – or von Wiese's four types of religious organizations – the ecclesia, the denomination, the sect and the cult[1]. Merton, in discussing the forms of interpersonal influence, lists the following types: coercion, domination, manipulation, clarification, provision of prototypes for imitation, advice, and exchange[2]. Kingsley Davis classifies social norms in traditional categories: folkways, mores, law, custom, morality, religion, convention, and fashion[3]. Lasswell sets up eight basic categories of values which he uses to classify institutions and leaders[4]; Malinowski sets up seven "basic needs" in terms of which cultural phenomena can be classified[5]; and so on.

---

[1] C. Wright Mills, White Collar, New York 1951, pp. 92 seq.

[2] l. c., pp. 144 seq.

[3] l. c., pp. 131 seq.

[4] l. c., pp. 174 seq.

[5] l. c., Introduction, p. xv.

[6] Katherine M. Wolfe and Marjorie Fiske, The Children Talk About Comics, in: Lazarsfeld and Stanton, l. c.

[7] For examples: Merton, Mass Persuasion, l. c., pp. 50 seq.

[8] Eduard Spranger, Types of Men, Halle 1928; for an attempt to develop measuring instruments for Spranger's concepts, see P. E. Vernon and G. W. Allport, A Test for Personal Values, in: Journal of Abnormal and Social Psychology, Vol. 26 (1931), pp. 231 seq.

All of the above mentioned sets of categories are of far greater generality than those which arise in the analysis of a single empirical study of limited scope. They are the result of attempts at general analysis of a wide range of situations. On the other hand, in their formal aspect, they are similar to the other forms of preliminary categories discussed earlier. Some of them are quite unsystematized; others include in their definition most of the elements required to set up a logical structure of basic attributes from which they could be derived, but this has not been explicitly done.

A special kind of descriptive system which might be mentioned under this heading consists of ordered categories, which are set up as developmental stages or degrees along a continuum. Thus Piaget distinguished the stages of development of children's attitudes toward the rules of conduct, from "moral realism" in which the letter of the rule is absolute to "autonomous rationality" in which blind acceptance "withdraws in favor of the idea of justice and of mutual service."[6] Scheler sets up seven categories of knowledge which he orders along the dimension of increasing "artificiality"; (1) myth and legend; (2) knowledge implicit in the natural folk-language; (3) religious knowledge; (4) the basic types of mystical knowledge; (5) philosophical-metaphysical sciences; (7) technological knowledge[7]. In such stets of categories an ordering along one dimension is explicitly stated, while the other attributes characterizing the categories are simply listed or suggested without any systematization.

---

[1] Leopold von Wiese and Howard Becker, Systematic Sociology, New York 1932, pp. 624 seq.

[2] Robert K. Merton, Patterns of Influence, in: Lazarsfeld and Stanton, l. c., p. 218.

[3] Kingsley Davis, Human Society, New York 1950, Ch. III.

[4] Harold D. Lasswell and Abraham Kaplan, Power and Society, New Haven 1950, pp. 55 seq.

[5] Bronislaw Malinowsky, The Group and the Individual in Functional Analysis, in: American Journal of Sociology, Vol. XLIV, No. 6 (May 1939), pp. 938 seq.

[6] Jean Piaget, The Moral Judgement of the Child, New York 1932.

[7] Max Scheler, Die Wissensformen und die Gesellschaft, Leipzig 1926, p. 62. The list given here is adapted from Merton's discussion of Scheler, in: Social Theory and Social Structure, l. c., pp. 230 seq.

Systematic Typologies

The most highly developed form of descriptive system which can arise in a qualitative analysis is one in which each type is explicitly derived from the logical combination of basic attributes or dimensions. A simple example is the logical scheme set up by Riesman in his study of political participation[1]. By examining a set of concrete "type cases" Riesman was led to break the concept of participation into two basic elements: emotional involvement and competence, or more simply, "caring" and "knowing". Taking each of these elements as a simple dichotomy, Riesman obtained four types of relations to politics:

		Competence +	Competence —
Affect	+	"involved"	"indignants"
	—	"inside-dopesters"	"indifferent"

Merton employs this technique in his typology of prejudice and discrimination. He starts from the usual formulation of two types of people: people who live up to the American creed of non-discrimination, and people who violate it. Merton suggests a further elaboration: that people be distinguished on one hand by whether they personally believe in the creed or not, and, on the other hand, by whether they practice discrimination or not.

> "... This is the salient consideration: conduct may or may not conform with individuals' own beliefs concerning the moral claims of all men to equal opportunity. Stated in formal sociological terms, this asserts that attitudes and overt behavior vary independently. Prejudicial attitudes need not coincide with discriminatory behavior. The implications of this statement can be drawn out in terms of a logical syntax whereby the variables are diversely combined, as can be seen in the following typology."[2]

---

[1] David Riesman and Nathan Glazer, Criteria for Political Apathy, in: Alvin W. Gouldner, Studies in Leadership, New York 1950, pp. 535 seq.

[2] Robert K. Merton, Discrimination and the American Creed, in: R. M. MacIver, Discrimination and National Welfare, New York 1949. — Table is adapted from Merton's, p. 103. See also Merton's Appendix to that paper: A Note on the Use of Paradigms in Qualitative Analysis.

Allen H. Barton and Paul F. Lazarsfeld

A Typology of Ethnic Prejudice and Discrimination

		ATTITUDE DIMENSION	
		Non-prejudiced	Prejudiced
BEHAVIOR DIMENSION	Non-discrimination	Type I: „The All-Weather Liberal"	Type III: „The Fair-Weather Illiberal"
	Discrimination	Type II: „The Fair-Weather Liberal"	Type IV: „The All-Weather Illiberal"

A mere list of the "folk-labels" of each type would appear superficially like one of the preliminary lists of categories; they are fundamentally different, however, since they are systematically derived from the cross-tabulation of two basic dimensions.

The most elaborate use of systematic typologies is found in Talcott Parsons' recent works[1]. Parsons sets forward five dichotomous attributes:

1. Affectivity – Affective neutrality;
2. Self-orientation – Collectivity orientation;
3. Universalism – Particularism;
4. Ascription – Achievement;
5. Specificity – Diffuseness.

By combining these five "pattern variables" Parsons has been able to construct general categories for describing social relations, cultural systems, and personality systems.

The process of constructing systematic typologies need only be briefly summarized here. The starting point is often a good preliminary set of categories. By examining them one derives a small number of attributes which seem to provide the basis for the distinctions made, and sets these attributes up as a multidimensional system (an "attribute-space"). This operation has been termed the "substruction" of an attribute space to a typology. One can then examine all of the logically possible combinations of the basic attributes. This serves to locate the original set of categories within the system; it often shows that some combinations have been ignored (appear as blank cells), while in other bases distinctions have been missed (the original category will overlap several cells). Of course not all of the logically possible combinations may be important or even empirically possible; it will often be necessary to restrict the combinations to be studied, or to recombine several categories to simplify the analysis. Such a recombination has been termed a "reduction", and is closely related to the operation of index formation[1].

---

[1] Talcott Parsons and Edward A. Shils, Toward a General Theory of Action, Cambridge, Mass., 1951, Part 2.

Partial Substructions

There remains to be mentioned a type of operation which is very frequent in qualitative analyses: the partial systematization of a concept or a set of categories. A good introduction to this operation is the well-known discussion by Simmel of envy and jealousy[2]. The situations in which these feelings arise are quite complex, and Simmel does not give an exhaustive account of them. What he does, however, is to indicate one important aspect in which the two attitudes differ: in the case of jealousy the person feels that he has a claim on the object of his desire, while in the case of envy he has no claim, only desire for the object. Simmel has thus partially substructed the attribute-space by which envy and jealousy could be systematically defined; he has not done so completely, but rather only enough to make one major distinction.

A more elaborate but still partial substruction is presented in Werner Landecker's discussion of "Types of Integration and Their Measurement". Landecker begins by indicating his discontent with the undifferentiated concept of "social integration". To study the relation of integration to other variables, to find its necessary conditions and its consequences, the broad abstraction must be broken down:

> "Early in the exploration of a type of phenomena it seems advisable to break it up into as many subtypes as one can distinguish and to use each subdivision as a variable for research. This appears to be a more fruitful procedure than to attempt immediately to generalize about the generic type as a whole. The main advantage of subclassification in an initial phase of research is that it leads to problems of relationship among subtypes which would evade the attention of the investigator if he were to deal with the broader type from the very beginning. Generalizations on the higher level of abstraction will suggest themselves as a matter of course once regularities common to several subtypes are discovered."[3]

In analyzing the concept of social integration, Landecker first breaks down society into two types of elements: cultural standards, on the one hand, and persons and their behavior, on the other. The logical interrelations among these two elements give him three types of integration:

---

[1] A more detailed discussion will be found in Paul F. Lazarsfeld and Allen H. Barton, Qualitative Measurement in the Social Sciences, in: Daniel Lerner and Harold D. Lasswell, The Policy Sciences, Stanford, California, 1951, pp. 169 seq.

[2] Georg Simmel, The Sociology of Conflict, in: American Journal of Sociology, Vol. IX, No. 4 (January 1904), pp. 521 seq.

[3] Werner S. Landecker, Types of Integration and Their Measurement, in: American Journal of Sociology, Vol. LVI, No. 4 (January 1951), p. 332.

"Cultural integration": integration within the realm of cultural standards;
"Normative integration": integration between cultural standards and the behavior of persons;
"Integration among persons": integration within the realm of behavior.

This last type in turn is broken down in terms of two types of human behavior: the interchange of ideas and the interchange of services. Integration within the realm of communication is termed "communicative integration"; integration within the realm of services is termed "functional integration".

Since "integration" is a relational concept, Landecker's types can be easily represented by a relational matrix – a table, along each side of which we list the elements involved in interrelationships. The interior cells of the table then indicate the logically possible connections, including, in the main diagonal, the internal relationship within each element:

	Cultural Standards	Persons and their behavior:	
		Communication	Work
Cultural Standards	1. Cultural integration	2. Normative integration	
Communication		3. Communicative integration	5. ?
Work			4. Functional integration

This relational scheme allows us to locate all of the types of integration proposed by Landecker. It also raises the question of further relationships not discussed and distinctions not made: e. g. the possible subdivision of "normative integration" in terms of the two spheres of behavior, the relationship between the two forms of behavior themselves, and indeed the possibility of distinguishing still other major spheres within the realm of behavior, for example the sphere of government, religion, or family life. For this reason we refer to Landecker's scheme as a "partial" substruction, one which is not fully worked through in all its logical possibilities.

An even more elaborate relational scheme is implicit in a discussion of "craftsmanship" by C. Wright Mills:

> "Craftsmanship as a fully idealized model or work gratification involves six major features: [1] There is no ulterior motive in work other than the product being made and the processes of its creation. [2] The details of daily work are meaningful because they are not detached in the worker's mind from the product of the work. [3] The worker is free to control his own working action. [4] The craftsman is thus able to learn from his work; and to use and develop his capacity

and skills in its prosecution. [5] There is no split of work and play, or work and culture. [6] The craftsman's way of livelihood determines and infuses his entire mode of living."[1]

In effect, Mills proposes six attributes by which a job situation can be described. If all six of these attributes have the values indicated above, we have the ideal-type situation of "craftsmanship". The situation of the modern industrial or office worker, Mills implies, is the opposite of the idealized craftsman in all these respects. Actually the six attributes give 64 logically possible combinations of values; the intermediate, mixed combinations however do not enter into Mills' present discussion, which deals only with the idealtype cases and not with the whole attribute-space.

In this case each of the six attributes actually refers to a relationship – between a worker's capacities and his work, between work and leisure, etc. They can be derived from a relational matrix consisting of four elements: the worker (his capacities, his character); the work activity; the final product; and the worker's leisure activities (his "play", "culture", "general mode of living"). Each of these can act on any of the others, as summarized in the relational scheme below:

	Worker	Work activity	Final product	Leisure
Acting on: Worker	—	(1) Gratifies (4) Develops	(1) Gratifies	
Work activity	(3) Freely controls	—		(5, 6) Contributes to
Final product		(2) Visibly related to	—	
Leisure		(5, 6) Contributes to		—

This scheme might suggest additional dimensions of the man-job relationship to be taken into account, to make for a more systematic classification of work situations. Still more elements might also be added – for instance, the "external rewards" which are not supposed to dominate the craftman's approach to his work but which are obviously primary for many kinds of jobs.

Formal devices such as attribute-space and relational matrices can often help to clarify concepts which are not systematically presented; sometimes they can even suggest significant possibilities not originally considered. They are not, however, a substitute for careful study and sensitive thinking about a problem. There are any number of possible relationships or attributes which might be picked out and put into a formal scheme; the strategic act is to "feel out" those which are important, which will ultimately help to solve the problems in which we are interested. Formal analysis can then be used to clarify, develop, and communicate the results of qualitative insights.

---

[1] Mills, l. c., p. 220.

III. QUALITATIVE DATA SUGGESTING RELATIONSHIPS

The only fully adequate way to test the existence of a relationship between two variables is through statistical analysis; to test cause-and-effect relations requires either a controlled experiment, or a rather large number of cases of "natural change" observed over time. But research which has neither statistical weight nor experimental design, research based only on qualitative descriptions of a small number of cases, can nonetheless play the important role of suggesting possible relationship, causes, effects, and even dynamic processes. Indeed, it can be argued that only research which provides a wealth of miscellaneous, unplanned impressions and observations can play this role. Those who try to get suggestions for possible explanatory factors for statistical results solely from looking at tabulations of the few variables which were deliberately included in the study in advance often can make no progress; sometimes even a single written-in comment by a respondent will provide a clue to additional factors.

Finding "Factors" Influencing Action

A classic case of the use of qualitative observation to disclose possible factors influencing behavior is the Western Electric study[1]. When the experimental group of workers maintained their high production even when physical conditions were made worse than before the experiment began, it was clear that something else was effecting their production. What the real factors were was first suggested by informal conversations with and observations of the experimental group, and from then on the main research effort was focussed on qualitative interviewing and observation to discover social factors and processes.

A recent study, the main focus of which was to uncover possible factors rather than strictly to test them, is Merton's Mass Persuasion[2]. Here some 75 people were selected who were known to have bought bonds through a Kate Smith war-bond "marathon" broadcast, and interviewed in such a way as to reconstruct their experience during the broadcast as spontaneously as possible. Among the factors which stood out as possibly influential were the fact that

---

[1] Roethlisberger and Dickson, l. c.

[2] Merton, Mass Persuasion, l. c.

"the all-day series of appeals emerged as a dramatic event . . . a single unified pattern"; "There was reciprocal interplay, for the audience was not only responding to Smith, but she was also responding to her audience and modifying her subsequent comments as a result"; "there was considerable qualitative evidence that belief in Smith's disinterestedness and altruism played an integral role in the process of persuasion"; "the audience's images of Smith, the class structure of our society, the cultural standards of distinct strata of the population, and socially induced expectations, feelings, tensions were all intricately involved in the patterns of response to the bond drive"; "but the cumulation of affect and emotion was not the major function of the marathon broadcasts. Above all, the presumed stress and strain of the eighteen-hour series of broadcasts served to validate Smith's sincerity . . . for an understanding of the process of persuasion, the most significant feature of these responses to the marathon is the effectiveness of this propaganda of the deed among the very people who were distrustful and skeptical of mere exhortation."

A study applying the same technique to a broader historical situation was that of Elizabeth Zerner on the factors in recent history which influenced attitudes toward Jews in France[1]. By a very small number of detailed interviews with people who were presumed to be good observers (about half concierges of apartment houses, the other half intellectuals), it was suggested that there were four main events influencing attitudes toward Jews in one manner or another: the persecutions outside France, which made people more aware of the Jews as a special group; the appearance of Jewish refugees from other countries, who were a clearly visible, different group in French society; the persecution of the French Jews during the occupation, which aroused certain feelings of guilt and a certain real danger for those who helped the Jews; the restoration of Jewish jobs and property after the Liberation, which obviously caused loss and disturbance to some non-Jews.

One special technique for discovering additional factors relevant to a given type of behavior is the examination of cases which deviate from the behavior expected in terms of known factors. Thus a purely economic-interest explanation of voting leaves poor Republicans and rich Democrats as "deviants"; by qualitative interviewing one may be able to get some idea of the factors other than economic interest which motivate voting. In the "Mr. Bigott" study of the response of prejudiced people to anti-prejudice cartoons, it was expected that the prejudiced would misunderstand the message of the cartoons – as indeed almost ²/₃ did. But there remained deviant cases who understood the

---

[1] Elizabeth Zerner (with Robert T. Bowers), German Occupation and Anti-Semitism in France, in: Public Opinion Quarterly, Vol. 11, No. 2 (Summer 1948), pp. 258 seq.

"hostile" meaning. To explain these cases, qualitative interviews were used. It appeared that such factors were involved as the degree of security in one's attitudes, the feeling that one's beliefs were socially caused and not a personal responsibility, the fact that the subject totally disidentified himself from the caricatured figure of "Mr. Bigott" and therefore were under no threat[1].

Qualitative Suggestions of Process

The simplest form of a "process" analysis is that which looks for an intervening variable which "explains" the correlation between two other variables. In his study of an East-coast slum neighborhood, William Whyte arrived (on a qualitative basis) at the following relationship: the socially aspiring "college boys" clubs seemed to be more unstable and subject to internal conflict than those of the none-mobile "corner-boys". To explain this relationship (which could be considered quite "upside-down" from a middle-class viewpoint) Whyte introduced a third variable, "informal organization". The corner boy clubs could draw for cohesion on already existing informal organization:

> "The daily activities of the corner boys determined the relative positions of members and allocated responsibilities and obligations within the group."

Among the college boys on the other hand,

> "Outside club meetings the members seldom associated together except in pairs. Since there was no informal organization to bind the men together, there was also no common understanding upon matters of authority, responsibility, and obligation."[2]

To be able to assert this explanation with any certainty, it would be necessary for Whyte to have observed corner boy clubs which were weak in informal organization and college boy groups which were strong in it; if the former were also unstable, while the latter were stable, it would constitute a certain test of the hypothesis. Whyte does not record whether he sought out such "test situations" or was familiar with a range of such cases.

In the same way the relation between membership in a corner gang and failure to rise economically was explained in terms of the impact of group relations on saving and spending habits. Whyte denies that preexisting differences in intelligence and ability explain the whole relation (again on the basis of qualitative observations which presumably held ability constant for a number of cases in each group). Whyte goes on to suggest:

---

[1] Kendall-Wolf, l. c., pp. [illegible] seq.

[2] Whyte, l. c., pp. 96 seq.

> "The pattern of social mobility in Cornerville can best be understood when it is contrasted with the pattern of corner-boy activity. One of the most important divergences arises in matters involving the expenditure of money. The college boys fit in with an economy of savings and investment. The corner boys fit in with a spending economy. The college boy must save his money in order to finance his education and launch his business or professional career. He therefore cultivates the middle-class virtue of thrift. In order to participate in group activities, the corner boy must share his money with others. If he has money and his friend does not, he is expected to do the spending for both of them . . . Prestige and influence depend in part upon free spending."[1]

This observation indicates some of the factors in the process of social mobility. Of course behind each such factor uncovered are other factors – the variables which for instance determine who sticks with the boys and spends, and who breaks away, saves money, and rises.

The uncovering of possible processes can go much further than inserting a third variable in a chain. The study of anti-prejudice cartoons mentioned previously suggested a whole chain which led up to misunderstanding among the non-deviant two thirds of the prejudiced people who misunderstood:

1. Identification with Mr. Bigott and momentary understanding;
2. Desire for escape from identification;
3. Disidentification mechanism (caricaturing Mr. Bigott, making him intellectually or socially inferior);
4. Derailment of understanding: absorbtion in the derogatory characteristics of Mr. Bigott to the exclusion of understanding the point of the cartoons[2].

In the Whyte study one finds the process of the rise of a local corner boy to political leadership traced out through a series of steps, with interacting forces noted[3]. To get a start, the corner boy must demonstrate his loyalty and ability to get results for his immediate circle of friends. Yet "if he concentrates on serving his own group, he will never win widespread support." "In order to win support he must deal with important people who influence other groups." Since he has only limited resources in terms of energy and access to official favors, he must "betray" his original friends by neglecting their interests and using his resources to help outsiders and big-shots. The result is a widespread cynicism toward "politicians" among the rank and file, which might be expected to cause constant turnover. However, according to Whyte, the politician is normally able to "trickle down" enough benefits to his followers in the district as a whole to prevent a revolt, even though his closest original friends who had the highest expectations may be badly disillusioned. The process reaches a kind of equilibrium, presumably at a level determined by the political abilities, initial "connections", and good luck of the individual politician.

---

[1] l. c., p. 106.

[2] Kendall-Wolfe, l. c., pp. 163 seq.

[3] Whyte, l. c., pp. 209 seq.

In exploring for possible factors affecting some given variable, or for chains of causes and effects constituting a "process", there appear to be two basic techniques. The first attempts to obtain objective information about the sequence of events, particularly what events preceded the response under investigation[1]. The typical questions, whether addressed to a subject or used by an observer to guide his observations, are: "What happened before X? What happened just before the subject made his decision to move, vote, buy, steal, etc.? What was the frame of mind? What had been going on in the family, neighborhood, nation, world? Had he been talking with anyone, reading anything, listening to anything?" Some responses will look like causal factors immediately, on the basis of our past experience or general hypotheses about human behavior. Others will only become prominent when we notice an apparent correlation between them and the criterion behavior in several cases.

The second technique is to ask people themselves to explain what happened and to give their reasons for acting as they did[2]. The basic question here is always "Why?" This technique has obvious limitations: people are often unaware of the real motives, of indirect influences, of the precise chain of causes and effects, of underlying necessary conditions. On the other hand it stands to reason that the participant knows a good deal about his own behavior, particularly about attitudes, motives, influences, "trigger events", and so on, and often can tell the outside investigator about things which he would never have guessed by himself. "Reasons" may not be the whole story, but they are an important source of information on possible factors, and in some cases a quite indispensible source, especially in the early stages of investigation. By adding to the general "why" query a set of more specific questions, focussing the respondent's attention on each of several basic aspects of the situation, reason questions can obtain more adequate coverage, although still limited to what the respondent himself is in a position to know.

Both of these techniques are combined in a technique of qualitative exploration of causal relations known as "discerning". This has been carefully described in Mirra Komarovsky's study of the effect of unemployment on the family status of the husband[3]. With only 59 case-studies to analyze, it was not possible to undertake a full-scale statistical analysis of the interrelations between all the possible variables. What was done was to take each case of apparent change due to the husband's unemployment and subject it to systematic checks: Had the change already begun before the unemployment? Did other factors arise concurrently with unemployment which might have been the real cause? Are the participants able to trace the step-by-step development of the change, the detailed links between unemployment and the altered role of the husband? If the respondents believe that unemployment was the reason for a certain change, on what evidence do they base their opinion? By these techniques it was possible to isolate with considerable promise of validity the causal relations between unemployment and family structure. The search for "possible factors" and "possible consequences" was made systematic; within the limitations of the data real precautions were taken against spurious relations.

---

[1] Robert K. Merton and Patricia L. Kendall, The Focussed Interview, in: American Journal of Sociology, Vol. LI (1946), pp. 541 seq.; Herta Herzog, Training Guide on the Techniques of Qualitative Interviewing, New York, Bureau of Applied Social Research, 1948; mimeographed.

[2] Paul F. Lazarsfeld, The Art of Asking Why, in: National Marketing Review, Vol. I, pp. 26 seq.

[3] Mirra Komarovsky, The Unemployed Man and His Family, New York 1940; especially pp. 135 seq.

Quasi-Statistics

Previous sections have dealt with operations of qualitative analysis which are essentially prior to quantitative research: observations which raise problems, the formulation of descriptive categories, the uncovering of possible causal factors or chains of causation for a particular piece of behavior. These operations stimulate and focus later quantitative research, and they set up the dimensions and categories along the "stub" of the tables, into which quantitative research may fill the actual frequencies and measurements.

However, one encounters very frequently in social science literature studies which do not use the mechanism of quantitative data-collection and statistical analysis, and still make the kind of statements which quantitative research makes. These statements may be simple frequency distributions (i. e. "most Trobrianders" or even "the Trobriander" knows or believes or does so-and-so); they may be correlations (corner boys have a spending economy, while college boys have a saving economy); they may be statements of causal relationships ("If [the politician] concentrates upon serving his own group, he will never win widespread support . . . In order to win support, he must deal with important people who influence other groups"). Such statements, based on a body of observations which are not formally tabulated and analyzed statistically, may be termed "quasi-statistics". They include "quasi-distributions", "quasi-correlations", and even "quasi-experimental data".

Non-quantitative research of this sort is no longer logically prior to statistical research. It rather directly substitutes for statistical research, making the same kind of statements but on the basis of a recording and analysis of cases which takes place largely within the mind of the observer. This kind of research has obvious shortcomings, but it also has a place in the research process, viewed as a continuing and increasingly refined pursuit by the whole community of social scientists.

An example of the dangers of impressionistic "quasi-statistics" is given by Bernard Barber in an article on participation in voluntary associations:

> "American observers themselves were overwhelmed by what they did not fully understand: instance the following from Charles and Mary Beard's *The Rise of American Civilization:* ,The Tendency of Americans to unite with their fellows for varied purposes ... now became a general mania ... It was a rare American who was not a member of four or five societies ... Any citizen who refused to affiliate with one or more associations became an object of curiosity, if not suspicion'. Although in comparative perspective the United States may well be a 'nation of joiners', a survey of the available data on the number of people with memberships in voluntary associations reveals the little-known fact that many have not even a single such affiliation. This uniformity too holds for all types of areas in the United States, whether urban, suburban, small city, small town or rural."[1]

Barber then quotes statistic showing that in these various areas and strata of the population from one third to over two thirds of the people do not belong to any voluntary associations. As one proceeds from simple frequency distributions to correlations and then to systems of dynamic relationships between several variables, impressionistic "quasi-statistics" become steadily less adequate.

On the other hand it is argued that a careful observer who is aware of the need to sample all groups in the population with which he is concerned, who is aware of the "visibility bias" of the spectacular as opposed to the unspectacular case, who becomes intimately familiar with his material over a long period of time through direct observation, will be able to approximate the results of statistical investigation, while avoiding the considerable expense and practical difficulty of quantitative investigation. It has been claimed, for instance, that to provide a fully statistical basis for the conclusions which Whyte was able to draw from his observation of corner gangs and college boys groups, would require hundreds of observers studying hundreds of gangs and neighborhoods over many years.

There are some situations in which formal quantitative methods are apparently less necessary than others. When one is dealing with primitive groups with a nearly homogeneous culture, in which one set of prescribed roles is just about universally carried out by the population, it may require only the observation and interviewing of relatively few cases to establish the whole pattern. The same argument can be applied to studies of a quite homogeneous sub-culture within a civilized society. These methods seem to have succeeded in presenting a good first approximation at least in the description of the culture and behavior of such groups. When anthropologists now call for formal sampling, data-recording, and statistical analysis it is either to catch up finer details – the small number of deviant individuals, for instance – or to deal with situations of culture groups with less homogeneity – with groups in process of acculturation, breakdown of old norms, or the development of strong internal differentiation.

---

[1] Bernard Barber, Participation and Mass Apathy in Associations, in: Gouldner, l.c., pp. 481 seq.

In situations of less homogeneity and simplicity, it is doubtful that quasi-statistics are anything like a full substitute for actual statistics. However they can still play an important "exploratory" function. Statistical research is too expensive and time-consuming to be applied on all fronts at once; like the 200-inch telescope it must focus on a few areas of particular interest for intensive study. Quasi-statistical studies can run ahead of the more cumbersome quantitative procedures to cover wide areas of social phenomena, and to probe into tangled complexes of relationship in search of possible "processes". They serve as a broad scanner and "finder" like the wideangled but less powerful Schmidt telescope of Mount Wilson and Palomar. Moreover the gathering and analysis of "quasi-statistical data" can probably be made more systematic than it has been in the past, if the logical structure of quantitative research at least is kept in mind to give general warnings and directions to the qualitative observer.

Systematic Comparison

There is one special form of research into relationship which stands on the border between statistical and quasi-statistical methods. This involves the systematic comparison of a relatively small number of cases. It differs from quasi-statistics in that the cases are proceeds along lines closely approximating those of a statistical survey or controlled experiment. However it involves too few cases to actually apply statistical tests, and it involves natural situations in which one cannot be certain that "other factors are equal" for the various cases beyond those factors specifically analyzed. It is as though one set up the tables for a statistical or experimental research, but had only one or two cases to fill in each cell, and perhaps had to leave some entirely empty.

This form of "comperative research" is the only possible when the "cases" to be studied are social phenomena of a high order of complexity, such as wars, revolutions, large-scale social systems, forms of government. There do not exist very many cases in recorded history of such phenomena. Toynbee faced this problem in his classic "comparative analysis" of "Great Civilisations" – there were only about twenty-one such civilisations, along with a number of abortive or arrested civilisations[1]. Weber faced the same situation in dealing with the role of religious systems in the development of society[2]. Besides the total number of available cases being small, each is a very large and complex unit which requires great time and effort to analyze. Even where there are a large number of cases, this factor may compel the researcher to restrict himself to the systematic comparison of a few. This situation arises in studying communities or large institutions. To describe any one community's social structure is such a large job that most studies have been of single cases. Only after different researchers over a generation have produced a dozen or so such studies can a "secondary analyst" undertake a comparative analysis. In the long run, it is to be hoped that data-gathering procedures on such complex "cases" can be so simplified that statistical studies will become possible. Then the intensive study of one community, factory, union, government agency, or voluntary association can give way to a quantitative study of a sample of such cases, testing the hypotheses derived from single-case studies. Of course, where the difficulty lies in the fact that there is only a handful of case t all, the comparative method ist the best we can do.

An example of systematic comparison of a small number of cases is offered by Lipset's study of the Canadian province of Saskatchewan. The population of the province, mainly wheat farmers, had a remarkably high level of participation in political affairs and in collection for this unusual behavior, comparisons were first made with areas where participation was known to be low.

The amount of participation in public affairs in the large cities of Canada and the U. S. is notoriously low. This is true even of cities like Toronto and Vancouver, which resemble Saskatchewan in giving a large vote to the new, radical C. C. F. party. Comparing Saskatchewan with these large cities, Lipset was struck by the smallness of political units in Saskatchewan and the large number of offices to be filled in each. The average rural municipality had fewer than 400 families, with over 50 elective posts on municipal council and school boards to be filled, while most large cities elect no more officials than that to represent their hundreds of thousands of families.

Besides the small size of Saskatchewan communities, they were relatively lacking in social stratification – almost everyone was a working farmer. In this respect too Saskatchewan is at the opposite pole from the cities, with their wide differences in incomes and their staffs of specialists for performing normal public services. In the cities the positions of responsibility which are available tend to be monopolized by upper-class people and professionals; in Saskatchewan school boards, telephone companies, marketing agencies, etc. had to be staffed by ordinary farmers, who thereby acquired organizational and political skills unknown to the average city dweller.

---

[1] Arnold Toynbee, A Study of History, Abridgement by D. C. Somervell, New York 1947.

[2] Max Weber, Gesammelte Aufsätze zur Religionssoziologie, Tübingen 1920-21. A brief outline is presented in: Talcott Parsons. The Structure of Social Action, Glencoe, Illinois, 1949. Ch. XV.

Certain rural areas also are highly stratified, and in such areas the rate of mass participation is also low:

> "Within the rural areas of the Southern States or in parts of California, where significant social and economic cleavage exists within the rural community, the wealthier and upperclass farmers are the formal community leaders, and the bulk of the poorer farmers are politically apathetic." (p. 202)

A third structural factor distinguishing Saskatchewan is its exposure to extreme economic fluctuations, due to the unstable price of its one main crop and to the recurrence of drought. In this it can be contrasted with its eastern neighbor, Manitoba, which has more diversified crops, stable markets, and more reliable weather. And it is notable that Manitoba today has much less community and political activity. Low participation is also found in the Maritime Provinces, where the farmers generally have a low standard of living, but do not experience the chronic alternation between wealth and poverty of the farmers of Saskatchewan.

Having isolated these possible sources of high participation by comparing the social structure of Saskatchewan with that of areas of low participation, we can now look for other areas which have equally favorable patterns of social characteristics. The neighboring wheat-belt areas of North Dakota and Alberta have virtually the same characteristics: small political units, little social stratification, and highly unstable economies. And both these areas have widespread community participation through local government and cooperatives, and a readiness to develop new political movements when confronted by economic crisis. The same structural characteristics were found in Manitoba in the 1890's, at which time the Manitoba agrarian political movement evoked widespread participation. When Manitoba's society changed through the development of a large urban center with an upper-class and specialized services, and through the diversification of agriculture which ended the complete dependence on the wheat crop, mass participation in politics fell off. Manitoba thus provides a natural "before-and-after" experiment.

Lipset notes that the same pattern of structural characteristics which exists in Saskatchewan can also be found in communities far removed from the specific conditions of the wheat belt. Merton studied an American industrial community, "Craftown", which was small in size, relatively unstratified, being mainly inhabited by workers, and faced with a series of pressing social and economic problems. This community was found to be much more politically active than neighboring urban areas; its widerspread political participation resembled that of the Saskatchewan farmers (pp. 303-304). The generalization of explanatory factors from "the wheat economy" to attributes applicable to any community obviously opens up a much wider range of cases for use in comparative analysis.

This comparative analysis of areas of high and low participation can be summarized in the following scheme:

Scheme of Factors Accounting for Political Participation

(Attributes in parantheses were not explicitly discussed in the comparisons)

Economic fluctuations	Small social units	Little stratification	Cases:	Participation
+	+	+	Saskatchewan North Dakota Alberta Manitoba, 1890—1910 "Craftown"	High " " " "
(+)	(+)	—	Rural South Rural California	Low "
(+)	—	—	Large cities in U.S. and Canada	"
—	(+)	(+)	Maritime Provinces	"
—	—	—	Manitoba today	"

In the study just discussed it appears that comparisons were used not only to suggest explanatory factors but also to offer supporting evidence, as a kind of quasi-experimental test. One of the most celebrated instances of such quasi-experimental tests is found in Malinowski's study of the use of magic in the Trobriand Islands[1]. Malinowski wanted to test the old theory that primitive man uses magic because of a childlike confusion of the real and the imaginary or because of some instinctive belief in the supernatural. He found the Trobrianders engaging in some activities – for instance, fishing within the lagoons – for which their technology was adequate to permit certain economic returns and personal safety. Other activities -- for instance fishing in the open sea – involved uncertainty of return and risks to life which could not be eliminated by available technological means. In the safe and certain activities, no magic was used; in the unsafe and uncertain ones, magic was used a great deal. This supported Malinowski's contention that magic was not a substitute for rational techniques, but a supplement to them when dealing with situations beyond the power of available rational technology, which created severe emotional strain.

---

[1] Bronislaw Malinowski: Magic, Science and Religion. A brief outline is presented in Talcott Parsons, l. c., Ch. XV.

Of course the use of comparisons of small numbers of cases as tests requires great caution; care must be taken to see that other significant factors are in fact equal, and that cases are selected in an unbiased manner.

IV. MATRIX FORMULATIONS

Sometimes the analysis of qualitative observations confronts a mass of particular facts of such great number and variety that it seems quite unworkable to treat them individually as descriptive attributes or in terms of their specific interrelationships. In such a situation the analyst will often come up with a descriptive concept on a higher level which manages to embrace and sum up a great wealth of particular observations in a single formula. Take for instance, Ruth Benedict's description of the Zuni Indians, which mentions their avoidance of drugs and alcohol, their lack of personal visions, their placid response to divorce, their "mild and ceremonious" relation to their gods, and so on. After presenting a great many such particular facts, Benedict is able to sum them up in a single formula: The Zuni culture has an Apollonian pattern – that is, a central theme of avoidance of emotional excess. This pattern or theme permeates every aspect of Zuni life[1]. Such a formula capable of summing up in a single descriptive concept a great wealth of particular observations may be called a matrix formulation. This definition covers the notion of a "Basic pattern" of a culture, a "theme", an "ethos", a "zeitgeist" or "mentality of the time", a "national character", and on the level of the individual person a "personality type".

Matrix formulations may be applied to complex units at any level. In a study of an unemployed village in Austria, the researchers made a collection of separate "surprising observations". Although they now had more time, the people read fewer library books. Although subject to economic suffering, their political activity decreased. Those totally unemployed showed less effort to look for work in other towns than those who still had some kind of work. The children from the unemployed village had more limited aspirations for jobs and for Christmas presents than children of employed people. The researchers faced all kinds of practical difficulties because people often came late or failed to appear altogether for interviews. People walked slowly, arrangements for definite appointments were hard to make, "nothing seemed to work any more in the village".

---

[1] Ruth Benedict, Patterns of Culture, New York 1946, Ch. IV.

Out of all these observations there finally arose the over-all characterization of the village as "The tired Community". This formula seemed clearly to express the characteristics which permeated every sphere of behavior: although the people had nothing to do, they acted tired -- they seemed to suffer from a kind of general paralysis of mental energies[1].

In a study of a particular group – people who had been designated as "influentials" in an American community -- Merton confronted the problem of explaining their diverse behaviors. Various classifications proved of no avail in accounting for the wide range of observations available. The particular behavior on which the research was focussed -- the reading of news magazines – remained unexplained. In trying to order the "welter of discrete impressions not closely related one to the others", the researchers finally came up with one general "theme" which distinguished the influentials: some were "cosmopolitan", primarily interested in the world outside the local community, while others were "local", primarily interested in the local community itself. Merton clearly indicates the typical function of such a matrix formulation when he declares:

> "All other differences between the local and cosmopolitan influences seem to stem from their difference in basic orientation ... The difference in basic orientation is bound up with a variety of other differences: (1) in the structures of social relations in which each type is implicated; (2) in the roads they have travelled to their present positions in the influence-structure; (3) in the utilization of their present status for the exercise of interpersonal influence; and (4) in their communications behavior."[2]

The bulk of the article is then taken up with an elaboration of this dual matrix formulation in terms of all the specific behaviors which fit into one or the other type of orientation, the local or the cosmopolitan.

Matrix formulations can thus vary in the level of the unit which they describe, from a whole culture to a community and to a status group within a community. They are used right down to the level of individual personalities, where, for instance, a great variety of particular behaviors will be summed up in the matrix formula of an "anal personality" or a "cerebrotonic temperament". At the personality level they are often referred to as "syndromes", a term arising out of the physiological level where it refers to just the same kind of complex of individual facts all of which can be summed up in one single formula.

Another way in which the matrix formulations can vary is in terms of the relations between the elements. The elements which went into Benedict's formulation of an Apollonian culture were all alike in terms of the variable "emotional tone" – their emotional tone was low in intensity. They all went together in the same sense that one can classify in a single group all regions with a very even temperature, or all people with a high blood pressure. In a matrix formulation such as Tönnies' "Gemeinschaft", the elements seem to be involved in causal relations and processes with one another[1]. The element of "reciprocal trust" for example can be considered as growing out of the element of "prolonged face-to-face association with the same people", as can a great many of the other characteristics of a "Gemeinschaft"-situation. In the matrix formulation of an "anal personality", the behavior characteristics are thought of as all arising out of a single basic factor, the fixation of the erotic development at a certain childish stage. In many of the culture-pattern formulations since Benedict's purely descriptive ones, we find the idea that all of the elements in the pattern are products of the pattern of child training – or even of a single element in the child training pattern. Some matrix formulations involve a mixture of descriptively related and causally related elements.

---

[1] Jahoda and Zeisel, l. c.

[2] Merton, Patterns of Influence, l. c., p. 191.

Yet another way in which matrix formulae can be differentiated could be called their "projective distance". The following examples should indicate what is meant by this dimension. In Merton's study of *Mass Persuasion*, it was found that a wide variety of remarks made by the Kate Smith devotees could be summed up in the notion of "submissiveness to the status quo"[2]. They believed that it was right for some to be poor and others rich, they accepted their position in the system of stratification, they rationalized that the rich had so many troubles that it was just as well to be poor. The more general descriptive concept follows very directly from the manifest content of the respondent's statements; they could almost have made the generalization themselves.

In Cantril's study of the *Psychology of Social Movements*, he confronted a collection of interviews and observations of people who join all sorts of marginal political cults like the Townsend groups, Moral Rearmament, the Coughlinites, and so on[3]. Out of the welter of characteristics there emerged the general notion that all of these people were suffering from a lack of orientation in the complexities of the modern world, a need for a frame of reference within the events of their lives and for world affairs could be understandable. The matrix formulation of "need for orientation" seemed to tie together a great many diverse forms of behavior, attitudes and beliefs on the part of the members of these groups. Now this matrix formulation is further removed from the manifest content of the material than was the formulation in the previous example. There is a greater gap between the formula and what the people actually said, in talking about troubles they had and how good it made them feel that their movement told them what was wrong with the world and how it could be corrected. Here the statements and observations collected by the field work are interpreted as projections of a somewhat complex psychological state, which very few of the respondents themselves could directly articulate.

---

[1] Ferdinand Tonnies, Gemeinschaft und Gesellschaft, Leipzig 1887. A brief summary is found in Parsons, l. c., pp. 686 seq.

[2] Merton, Mass Persuasion, l. c., pp. 152 seq.

[3] Hadley Cantril, The Psychology of Social Movements, New York 1941.

A still greater distance between the manifest content of the material and the matrix formulation which is constructed to express its basic pattern is often found in the characterization of personality types or of the ethos of a culture. In these cases, guided by general theoretical orientations, one may use subtle indicators as a basis for a formulation which appears in some ways contradictory to the manifest content of the material. Verbal expressions and actual behavior patterns apparently indicative of feelings of superiority are interpreted, when seen in the context of more subtle indicators, as evidence of quite the opposite basic outlook. In Benedict's characterization of the Zuni as basically Apollonian, she has to explain away – on the basis of looking beneath the surface – various apparently "Dionysian" elements[1].

There is a good deal of similarity between the rationale of the matrix formulation and Parson's discussion of the place of "secondary descriptive systems" in his scheme of social systems based on the unit act:

> "When a certain degree of complexity is reached, however, to describe the system in full in terms of the action scheme would involve a degree of elaboration of details which would be very laborious and pedantic to work out. This is true even if description is limited to 'typical' unit acts and all the complex detailed variations of the completely concrete acts are passed over. Fortunately, as certain degrees of complexity are reached, there emerge other ways of describing the facts, the employment of which constitutes a convenient 'shorthand' that is adequate for a large number of scientific purposes.
>
> "... It has been seen that the acts and action systems of different individuals, in so far as they are mutually oriented in one another, constitute social relationships. In so far this interaction of the action systems of individuals is continuous and regular these relationships acquire certain identifiable, relatively constant properties or descriptive aspects ... It is not necessary to observe all the acts of the parties to a relationship, or all their attitudes, etc., but only enough to establish what is for the purpose in hand the relevant 'character' of the relationship ...
>
> "Thus the primary function of such a secondary descriptive scheme as that of social relationship is one of scientific economy, of reducing the amount of labor of observation and verification required before adequate judgements may be arrived at. A second function ... (is) to state the facts in a way that will prevent carrying unit analysis to a point where it would destroy relevant emergent properties."[2]

---

[1] Benedict, l. c., see p. 82, 83, 85, 94, 127, 112 for instances.

[2] Parsons, l. c., pp. 743 seq.

Parsons gives as examples of secondary descriptive schemes typologies of social relationships, personality types, and descriptive categories applied to groups. While Parsons' concept is not entirely identical with that of the matrix formulation, it illustrates much of the reasoning behind such complex descriptive concepts, which sum up and render manageable a large and varied body of individual points of data.

V. QUALITATIVE SUPPORT OF THEORY

So far we have mainly discussed ways in which qualitative data can contribute to the formulation of problems, classifications, and hypotheses. Qualitative materials are particularly suitable for this exploratory phase of research: their wealth of detailed descriptive elements gives the analyst the maximum opportunity to find clues and suggestions. For testing hypotheses, on the other hand, the ideal model would be the controlled experiment, with precise measurements on a limited number of preselected variables.

The use of controlled experiments in social science is increasing, but it remains severely limited. Recent years have also seen a great development of quantitative research, employing such rough approximations of the experimental design as the controlled observation of natural processes, or the correlational analysis of cross-sectional surveys. These techniques provide tests for certain theories. There remain, however, major areas in which theories are supported mainly by qualitative data.

The General Problem of Qualitative Support

The word "theory" has actually a number of different meanings, ranging from broad general orientations to precise propositions[1]. The theories for which qualitative support is most often used are relatively large-scale, wide-ranging systems or relationships. For example, large-scale theories of social change must rely upon the qualitative data of historical records; theories of the functioning of organizations and institutions are based largely on qualitative descriptions; theories of personality development grow out of clinical case materials. One calls to mind at once the historical theories of Marx, Weber's institutional analysis, Freud's personality theory, and the later work in their traditions.

In discussing the use of qualitative data to support theory, it should be made clear that more is involved than mere illustration. For illustration, intended to help the writer communicate the meaning of his concepts, purely imaginary examples can be used. Durkheim, for instance, in describing types of suicide, drew on examples from literature as well as real case-histories[1]. But for the use we have in mind it is important that the examples are real. Since they are not systematically samples or precisely measured, they do not offer rigorous proof in any statistical or experimental sense. Yet according to their number, range, and relation to the reader's own experience they offer varying degrees of support or corroboration. It is this function which we wish to examine.

---

[1] Merton, Social Theory and Social Structure, l. c., Chapter II.

Psychoanalysis, Marxism, and other theories of history in general or of overall personality development are large and complicated structures. The use of qualitative data in supporting such complex theoretical systems is one of the major undeveloped areas for methodological analysis. In order to find manageable illustrations of the use of qualitative data to support theory, we will restrict ourselves to a much simpler type, which may be called "trend theories". These are theories which call attention to one particular trend in society, usually derived from some underlying change in the economic or demographic structure.

We will consider three such theories. Erich Fromm has suggested that a major tendency on our time is "self-alienation", resulting from the insecurities and disruption of social bonds brought about by the rise of the market economy and industrialization. C. Wright Mills suggests that the rise of the big city and the standardization of tastes subjects increasing numbers of people to status-insecurity. Lasswell proposes the "developmental construct" of the garrison state, which attempts to work out the logical implications of the tendency toward an increasing reliance on military force in international relations.

"Signs of the Times"; Qualitative Observations Supporting Trend Theories

Fromm's theory as a whole is relatively complicated, dealing with the interaction of economic structure, personality, and systems of belief[2]. We will consider one particular aspect of that theory: that the individual, rendered isolated and powerless in the face of impersonal market forces, monopolies, mass organization and recurring wars, develops "automaton conformity" as a mechanism of escape:

> "This particular mechanism is the solution that the majority of normal individuals find in modern society. To put it briefly, the individual ceases to be himself; he adopts entirely the kind of personality offered to him by cultural patterns; and he therefore becomes exactly as all others are and as they expect him to be."[3]

---

[1] Emile Durkheim. Suicide. Glencoe, Illinois, 1951; Book Two, Ch. 6.

[2] Erich Fromm, Escape From Freedom, New York 1941; Appendix: Character and the Social Process.

[3] l. c., pp. 185 seq.

Let us see how Fromm goes about supporting this contention derived from his trend theory. First of all, he wants to demonstrate that it is at all possible for a person to think thoughts and feel feelings which are not his own, but induced from outside. This he does by describing a common hypnotic experiment where it is suggested to a subject that he will do certain things and have certain feelings upon awakening from the hypnotic sleep. This establishes that the phenomenon can exist; however it involves very special conditions.

Fromm then attempts to show that the same kind of behavior occurs in the daily life of many individuals. To specify what this would involve, he gives a hypothetical example of "pseudo-thinking" in daily life: the man who makes a weather prediction which he believes to be his own thinking, when he is simply repeating what he heard on the radio. He then proceeds to actual observations:

> "Many persons looking at a famous bit of scenery actually reproduce the pictures they have seen of it numerous times, say on postal cards . . . Or, in experiencing an accident which occurs in their presence, they see or hear the situation in terms of the newspaper report they anticipate."
>
> "The average person who goes to a museum and looks at a picture by a famous painter, say Rembrandt, judges it to be a beautiful and impressive picture. If we analyze his judgement, we find that he does not have any particular inner response to the picture but thinks it is beautiful because he knows that he is supposed to think it beautiful."[1]

Fromm gives a number of other observations of pseudo-thinking and feeling: the man whose face goes solemn after he leaves a party where he was "gay", children who say they "like" to go to school every day, people who believe they are marrying because they want to, while they are only conforming to other people's expectations, the case of a man who "voluntarily" follows the career set by his father, but suffers strange difficulties.

Of course Fromm's examples do not provide rigorous proof of the assertion that "the majority of normal individuals" are self-alienated. What Fromm has done is, first, to establish thoughts and feelings. Then he presented a hypothetical example to show how this might happen in daily life. Finally he mobilized a wide range of common observations to show that it actually does occour frequently.

Let us take another example. C. Wright Mills holds that the white-collar worker in urban, mass-production society suffers from an increasing ambiguity and insecurity of social status; he or she is unable to develop a stable self-esteem or a secure prestige-status in the eyes of others[2]. Mills presents qualitative observations from various sources. Strategically-placed specialists – e. g. personnel directors – report such surprising behavior as the following: a girl typing in a large office was deeply hurt when her chair was replaced with a new one, because the old one had her nameplate on the bottom; office workers attach great importance to objectively meaningless changes in the location of desks, etc. This suggested to Mills the intensity of the effort to retain some identity and status in a standardized environment.

---

[1] l. c., pp. 192 seq.

[2] Mills, l. c., Ch. 11, The Status Panic.

Detailed interviews found other phenomena which fitted in with the notion of a "status panic": Routine clerical workers tried to conceal the nature of their own work, and borrow prestige from the firm or industry, by identifying themselves with such phrases as "I am with Saks", or "I work at Time". They saved up their salaries and spent them for an evening at expensive places of entertainment, or for a vacation at a costly resort, in order to "buy a feeling, even if only for a short time, of higher status". A salesgirl dealing with "Park Avenue" customers will try to behave with greater dignity and distinction in her off-the-job contacts than the girl who works on 34th Street.

Such qualitative observations provide support for a theory in several ways. Aside from their own weight, they may call the reader's attention to certain areas of his own experience, which may provide much additional support. Furthermore, when the observations cover many different areas of behavior they gain additional weight, because they indicate that the theory has the ability to account for a wide range of phenomena.

The notion of the "garrison state" as formulated by Lasswell also suggested a trend growing out of basic social changes; however it was unusual in being formulated not ex post facto but at a very early stage in the process. This is what Lasswell meant in describing it as a "developmental construct" – a trend theory referring not only to the immediate situation but to the future. The substance of Lasswell's theory was a set of consequences which could be expected if the world situation stimulated modern great powers, with modern economies and technology, to all-out development of military strength. Some of these consequences were relatively logical and obvious: the entire labor force would be put to work, political conflicts will be minimized in the interests of national unity, the government will become involved in more and more previously private activities. Others were more of a symbolic nature:

> "The distinctive frame of reference in a fighting society is fighting effectiveness. All social change is translated into battle potential.
>
> "The military are therefore compelled to become competent in the skills of technology, administration, and public relations; there is a 'merging of skill' between the professional soldier and the manager of large-scale civilian enterprise.
>
> "There is intense concern with public morale and its manipulation, the military

'are compelled to consider the entire gamut of problems that arise in living together under modern conditions.'"[1]

If one wanted in 1953 to see to what extent the predicted trend had become a reality, the developmental construct provides a guiding framework for observation. "All social change is translated into battle potential" -- a study of illiteracy in America is discussed mainly in terms of "how many divisions" it costs the army. "The merging of skills" -- generals appear on boards of directors of large corporations, and large corporation executives appear in the defense departments. "Intense concern with public morale" -- the traditional concern with the morality of school-teachers in terms of smoking, drinking, and sex shifts to a concern with treasonable ideas. In politics the traditional charges of corruption and inefficiency against opponents are replaced by charges of treason, or of failure to build enough air groups.

One or two of these might be isolated events; taken together they begin to build up a "pattern", giving some plausibility to the theory. At this point it becomes important to undertake more systematic studies of the actual extent and degree of militarization, to distinguish its progress in different spheres, etc.

From the few examples presented here, it is hoped that the reader has been able to get an idea of the intermediate role played by qualitative observations in relation to theory -- as more than simply illustration, but less than definitive proof. It is likely that there are several degrees or stages of qualitative support, ranging from an initial encouragement to go on with a certain line of speculation, to a systematic examination of case material which offers some approximation to the classical canons of proof. There also may be differences in function according or the type of "theory" involved. The present analysis is only a tentative beginning; there is a great need for intensive work on additional examples of qualitative support of theory[2].

Now that we are at the end of our survey, it is necessary to add that some of the areas we have distinguished inevitably overlap at their borders. It would be easy to find examples in which it would be difficult to decide whether they have the complexity of a matrix formula. Another set of uncertainties will sometimes arise when we have to decide whether the writer is arguing for a relationship between a few variables or whether he is supporting a rather general theory. Even the very notion of qualitative research has its haziness on the fringes; it would not be easy, e. g., to say when the comparative analysis of a few cases shades over into statistical treatment.

---

[1] Harold Lasswell, The Garrison State and Specialists on Violence, in: American Journal of Sociology (January 1941). Reprinted in: Lasswell, The Analysis of Political Behavior, London 1947, pp. 146 seq.

[2] Paul F. Lazarsfeld, Remarks on Administrative and Critical Communications Research, in: Studies in Philosophy and Social Science, Vol. IX (May 1941) gives a number of further examples in the field of mass communications, especially from the work of Leo Lowenthal and T. W. Adorno. In another field, Thorstein Veblen: Theory of the Leisure Class offers a classic example of the use of qualitative observations.

The present discussion is a beginning only. It has started from a simple position: that there exists this area of research which is generally considered important but which has not been analyzed methodologically. It has set forward a collection of examples both as an extensional definition of "qualitative analysis" and as material for further study. Besides collecting this material, it has made a preliminary organization of it. There are many problems left unsolved in the discussion[1]. The tentative classifications set forward here need to be tried out on additional materials; more good cases need to be collected and examined. Only after many successive phases of logical formulation and attempted application will the methodology of qualitative techniques come to posses the same usefulness to the research worker which is today possessed by quantitative methodology.

---

[1] The authors are indebted to Professor Merton and the members of his seminar in the Sociology of Occupation for raising many of these problems in discussing an earlier draft of this paper.

INTRODUCING ANOTHER VARIABLE

One of the most important tasks in the analysis of data is to discover what we can about cause and effect relationships. This does not mean that we can conclusively prove our assertions, since they often turn on assumptions which have not yet been tested or which are untestable for all practical purposes. But we still must be as explicit as we can, showing the logical structure of our arguments and the points at which empirical data are relevant. It is the nature of this task which is discussed by Professors Goode and Hatt, using terms such as intervening variable, antecedent variable, specification, and so on, which are part of the working vocabulary of the social sciences.

From William J. Goode and Paul K. Hatt, Methods in Social Research, Copyright 1963 by McGraw-Hill Book Company. Used by permission of McGraw-Hill Book Company. pp. 313-358.

William J. Goode and Paul K. Hatt, Methods in Social Research (New York: McGraw-Hill, 1952), pp 313-358.

A main thread of thought running through this volume is that sociology rests upon the same foundations as do all other sciences. Its assumptions regarding the empirical world are the same, and the precautions which the sociologist must observe in carrying out research must also be observed by scientists in other fields. We are, then, at the beginning of an era in which the conscious application of scientific methods to sociological problems should be marked by great advances.

This promise is also based upon a fundamental assumption of sociology, constantly borne out by daily observation, that there is an orderliness within social phenomena. Just as the psychiatrist finds a system or structure within the seemingly most incoherent expressions of a schizophrene, so can the sociologist find predictability and order within the most anomic or disorganized strata and social groups. The intimate spontaneity of friendship groups exhibits this orderliness no less than the impersonal memoranda within a bureaucracy. Indeed, if the behavior of other members of society were not predictable, it would not be possible to communicate, interact, or maintain any of the common enterprises in which we take part.

It follows, then, that modern research must reject as a false dichotomy the separation between "qualitative" and "quantitative" studies, or between the "statistical" and the "nonstatistical" approach. The application of mathematics to sociology does not ensure rigor of proof, any more than the use of "insight" guarantees the significance of the research.

The fundamental questions to ask about all research techniques are those dealing with the precision, reliability, and relevance of the data and their analysis: (1) how precise are the observations? (2) can other scientists repeat the observations? and (3) do the data actually satisfy the demands of the problem, that is, do they actually demonstrate the conclusion? If the observations are crude, casting them in a statistical form will not help the research. If other scientists cannot repeat them, mathematical manipulation is futile. If the data do not satisfy a rigorous logic of proof, the conclusion remains doubtful.

Furthermore, no matter how precise measurement may be, *that which* is measured remains a *quality*. Quantification simply achieves greater precision and reliability in *measuring the qualities which are considered important*—intensity of anti-Semitic attitudes, degree of social cohesion, conformity with moral rules, etc. The process of achieving precision leads to the clarification of ideas and helps to recast substantive knowledge, but in a fundamental sense the research may nevertheless be called qualitative.

Similarly, the most "qualitative" of social research attempts rough *measurement*. The historian speaks of "a growing antimonarchical feeling during the reign of Louis XVI." The anthropologist contrasts the intensity of emotion aroused by, say, the murder of a kinsman as against the murder of a tribal enemy. The economist may single out those who find the security of income far more important than the promise of great rewards when they are considering the choice of an occupation. The sociologist comments that the individual who is well integrated within a strongly cohesive group is better able to withstand emotional shocks than other individuals. In each of these statements, there are implied *measurements* of important qualities. We may or may not be satisfied with the *degree* of precision in these notions, but they are essentially attempts to measure the effect of different variables. It is equally clear that, when data *are sufficiently precise*, statistical techniques can simplify the task of understanding them. Such techniques are aids in research, just as good methods of recording data can be, and they should be used whenever the problem permits it.

Thus, the increasing use of statistics is not the distinguishing feature of modern social research. Rather, it is the increasing precision and reliability of research techniques, and higher standards of proof, which have made the use of statistics more fruitful. In turn, the increasing *fruitfulness* of statistical manipulation has stimulated further developments in both sampling and statistical theory; while the *needs* of modern statistics have stimulated still greater precision and reliability in the collection of data.

These developments have occurred on many fronts, such as the better identification of the important social variables, increased precision of the questions used in schedules, and a better grasp of interviewing techniques, among others.

However, most of this growth may be classified as techniques for stimulating responses, or obtaining observations, *which are easier to categorize*. More precise questions allow the answers to be analyzed more easily. The mastery of interviewing permits deeper probing in the search for more precise answers from respondents. What, then, of those sources of data which have *not* been structured previously by such techniques? Most of our daily observation and experience, the newspapers and magazines we read, the radio programs to which we listen, as well as historical records and the recorded protocols of psychiatric and other depth interviews, are all essentially "unstructured" but may be important sources of data for certain sociological problems. If such data are eliminated from consideration, the range of information available is narrowed and much of the richness of social experience may be lost. On the other hand, if they are utilized as they occur, little order appears in them, and few fellow scientists might agree to any one interpretation of them.

Techniques are therefore being developed which permit us to order and analyze such data. Since most of these procedures are really ways of *classifying data which were not originally created for research purposes,* the term *qualitative coding* is usually applied to them. When qualitative coding is applied to the content of various communication media such as magazines, newspapers, radio programs, or similar materials, it is called *content analysis.* Since the most effective application of the *case history* or *case study* to social research depends in large part upon qualitative coding, that special problem is included in the following discussion.

Succeeding sections will, then, deal with these subjects: (1) simple coding operations, (2) qualitative coding, (3) content analysis as one application of qualitative coding, and (4) the case study.

SIMPLE CODING OPERATIONS

The student may feel that there is little point to a discussion of coding operations, since they are usually bracketed with large-scale surveys. Although it is true that such surveys almost always code the materials gathered, the student may find it profitable to consider whether a small project might also benefit from coding the data. If the class attempts a joint project, or breaks into several project groups, coding may be the most effective means of handling the data. A brief explanation of this tool will allow the student to make a decision concerning the use of coding.

When to code. Coding is an operation by which data are organized into classes, and a number or symbol is given to each item, according to the class in which it falls. Thus, counting the symbols gives us the total number of items in any given class. The basic operation, of course, is that of *classification.* Assigning the number or symbol to a given datum then becomes a mechanical procedure. *How* to classify must, naturally, depend upon the questions which have been asked and the concepts which are used in the particular research. These problems are discussed in the earlier chapters of this volume, and examples are given in this chapter. Let us now ask the practical question, "When is it profitable to code?"

The answer depends mainly on three variables: (1) the number of *respondents* or *sources* of data in our study; (2) the number of *questions* asked; and (3) the number and complexity of *statistical operations* planned for the study. If the number of cases is large, any kind of tabulation becomes difficult unless the data are coded. The individual sheets for each case become separated after a few shufflings, or they become torn and worn.

Small errors in counting the answers require handling all the sheets once more. By the use of coding procedures, however, retabulations may be avoided or minimized. If the number of questions is large, the same set of considerations holds.

Most important, however, there is no easy way to carry out complex cross tabulations without some form of coding. Any statistical operation requires the manipulation of numbers, which in turn must represent the data from the schedule. It is possible either to make many piles of sheets for each comparison or analysis, or to give numbers to the answers and summarize these separately on other sheets or cards. In the latter case, the operations are much simpler. The more complex the operations planned, the more useful is some form of coding.

At what stage to code. Coding can be carried out at any phase in the study, from the interview itself to the period just prior to the tabulations. In the section on formulating the questionnaire, mention was made of *precoded questions.* These were questions which had already been field-tested for meaning and range of possible answers, and which were physically set up on the schedule so that checking the answer automatically coded the data. An example would be the following:

	1	2
Are you a veteran of World War II? (Circle answer)	YES	NO

Answers which are set up in this fashion can be tabulated very easily by hand, or they may be punched directly onto cards for machine tabulation. In this case the interviewer is actually *coding* as he goes along, although no separate operation is required.

Similarly, the interviewer may be asked to do the coding as soon as he hears the answer. This can be done in a fairly cautious fashion, as an almost automatic operation:

How did your husband feel about this? 1. Strongly approved; 2. Mildly approved; 3. Indifferent; 4. Mildly disapproved; 5. Strongly disapproved. 2

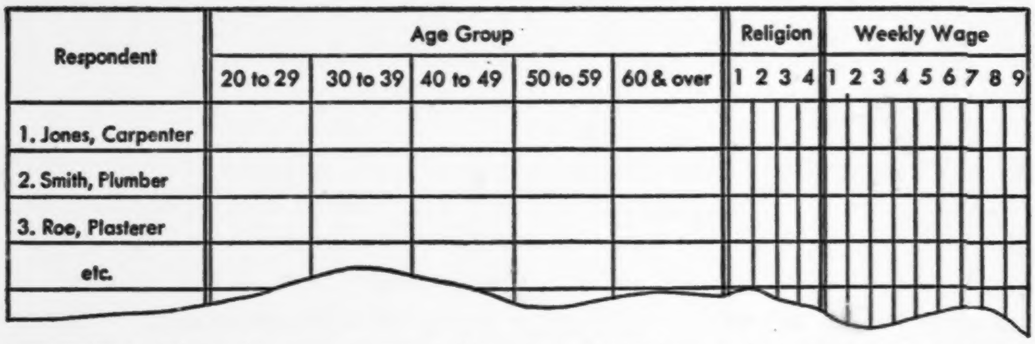
When the respondent answers, the interviewer has only to mark the proper number in the coding margin.

Suppose, however, that the goal is to classify respondents in terms of *annual income,* while many of them are laborers who typically think of their earnings in terms of daily wages. It would be possible to ask them for their total annual earnings directly. This, however, might cause them to make important errors in arithmetic. Second, the interviewer might carry out the operation mentally, coding the total earnings in its proper class after calculating daily earnings times the total estimated number of days or weeks worked. Although this is a simple procedure, experience shows that even the best interviewers may make errors. Most important, however, there is no way of discovering the error, since the only figure recorded will be the code number. Consequently, a third procedure is indicated: (1) recording the daily wage; (2) recording the estimated number of days or weeks worked; and (3) carrying out the necessary calculations and coding in the office, from these original figures. Although errors can be made, at least they can be checked.

The decision as to the *time* of coding must, then, be a matter of choosing the phase at which the least number of errors, and the greatest amount of control, will be achieved.

Paper aids to coding. When only simple tabulations are required, there is little advantage to *transcribing* the coding. A heavy colored pencil can be used to mark the code symbols opposite the answers, and the answers can be hand tabulated. If the number of cases is small, some type of *summary sheet* may be used. This may be attached to each schedule, for later independent use, or it may be designed on a larger scale to take care of *all* the cases. If a *special* sheet is used for *each* schedule, it should be very heavy so as to withstand considerable handling, and designed for ease of tabulation. Such sheets have the advantage that several persons can work on this stage of the operation at one time.

The *larger summary sheets,* containing data from *all* the cases, are sometimes useful for quick overviews of the data, especially when the number of cases is small. The accompanying illustration shows a simple sheet of this kind. The code symbols can be used instead of full headings, as a solution for the space problem. It is possible to design such sheets from accounting paper, affording an easy method for the cross tabulation of a small number of cases.



Respondent	Age Group					Religion				Weekly Wage								
	20 to 29	30 to 39	40 to 49	50 to 59	60 & over	1	2	3	4	1	2	3	4	5	6	7	8	9
1. Jones, Carpenter																		
2. Smith, Plumber																		
3. Roe, Plasterer																		
etc.																		

FIGURE 1. Example of a summary sheet for simplifying analysis of the study data.

However, when the numbers of cases go much beyond 150, the student may find some form of *punch card* of advantage. Unless IBM machines are easily available, the use of Hollerith cards is not a good solution for students. A great amount of time may be wasted in obtaining access to the machines, or in planning carefully the exact operations to be carried out, so that adequate written instructions to the operators can be given. The net result may be a loss in efficiency.

On the other hand, even for student group projects there may be a net advantage in using the McBee Keysort punch card. For both this card and the Hollerith card, it is necessary to use *numbers,* for the coding techniques are the same. The McBee card is relatively inexpensive and is procurable in several styles, depending on the number of items being coded. It is basically a card with a row of holes punched around its edge. In most studies, a single card will carry the data for one respondent or group. However, more than one card may be used, if an identifying punch or written symbol is used for both cards to prevent any confusion.

Each hole around the edge of the card is numbered. The technique, then, is to *open up* the hole corresponding to the code number. Thus, with a simple three-choice code, "Approve" (1), "Indifferent" (2), and "Disapprove" (3), and a response of "Approve," which is No. 1 in the code, the hole corresponding to No. 1, on the right-hand upper edge of the card, is opened as in the accompanying illustration. When this set of answers for all respondents has been thus coded and punched, it is possible to separate those in any desired category by inserting a needle through the holes. When the needle is lifted, all the cases corresponding to that punch will *drop out* (because that section of the card has been opened). The remaining cards, with intact holes, will remain on the needle. Having coded and punched *all* the answers in a similar fashion, it is possible to select all the males who are over twenty years of age, and who are married, by three insertions of the sorting needle. Since this method is used only on *small* samples, the actual counting is a minor task. What is important is that all the data have been classified and recorded permanently on the cards, for any kinds of complex operations which may later be decided upon.

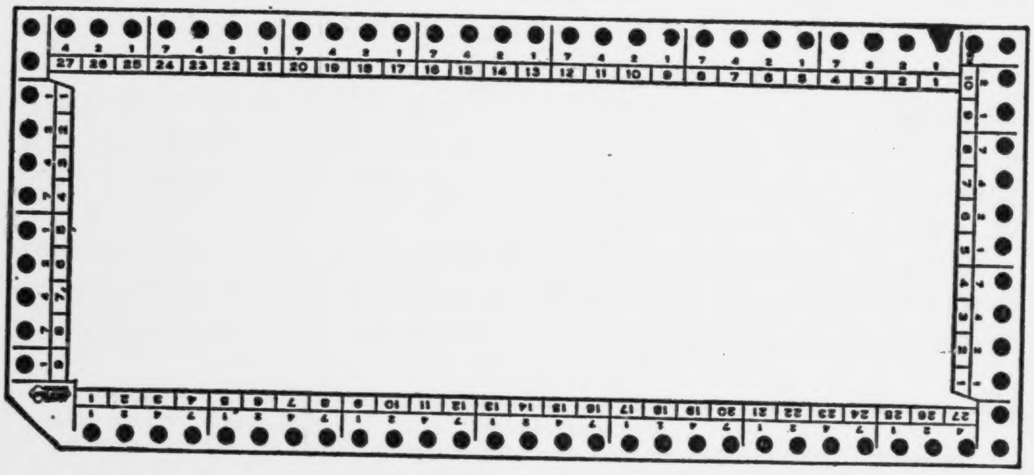


FIGURE 2. Sample of the McBee card, showing method of opening holes.

Since the number of questions may rapidly exhaust the number of holes, it is possible to use *combinations* of the numbers in the two rows. This requires *double* punching and thus a double insertion, but more data can be placed on the single card. Thus, for the first section in the upper right-hand corner, we could obtain eight classes, as shown in the table. Further

Coded as Hole Number

Top Edge of Card

Income Level		
$ 10.00 to $ 24.99	(punch)	1
25.00 to 39.99		2
40.00 to 54.99	(2 plus 1)	3
55.00 to 69.99		4
70.00 to 84.99	(4 plus 1)	5
85.00 to 99.99	(4 plus 2)	6
100.00 to 114.99		7
115.00 to 129.99	(7 plus 1)	8

instructions may be obtained from the McBee Company for still other operations made possible by the Keysort system.[1]

For large surveys, however, the need for greater ease and speed of sorting and tabulation requires the use of IBM and similar equipment. First developed for the needs of the United States census, such mechanical aids have come to be applied to many complex statistical problems in both business and research. The punch card used, which is called a Hollerith card, is reproduced here. Each one of its 80 columns contains 10 numerical choices, so that for any given item (such as "wages") it is possible to have 10 coding classes. There are two additional punches, called "fields," which can be used for further classes. However, these "fields" are more often used to direct the various machines to carry out statistical operations. That is, fairly complex statistical manipulations can be wired into the

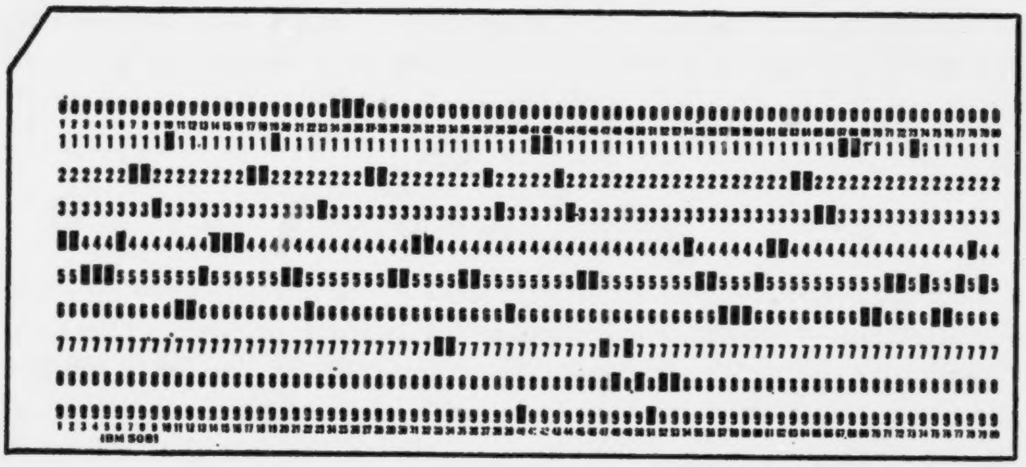


FIGURE 3. Sample of a punched Hollerith card used in IBM counting, sorting, and tabulating machinery.

---

[1] McBee Keysort Company, Detroit, Mich.

machine by changing the circuits on a removable wiring board. The "field punches," then, control some of these operations.

The advantages of such an aid are obvious: greater amount of information on each card, greater ease and speed in carrying out simple sorting and tabulation, and the possibility of performing complex statistical computations within the machines themselves. However, for most student projects and for most projects with low funds and small numbers of cases, these advantages do not outweigh the low cost and ease of independent operations represented by the Keysort system.

Checking the coding operations. Even when the student himself is carrying out the coding of his schedules, careful instructions should be written and should be followed without deviation. Further, such directions should be tested by having fellow students code a sample of the schedules according to the instructions. If the coding carried out by others deviates from his own, the instructions may be ambiguous, or they may fail to take into account some of the problems presented by the data. This precaution may uncover some difficulties in *classification* which the researcher has not clarified adequately.

Since it is possible to make errors no matter how simple the problems of classifying and coding, a definite sampling of the coding should be made, and every *n*th schedule should be recorded to see whether there are many errors. This is called "spot checking," and it is a further precaution against errors in the mechanics of coding.

QUALITATIVE CODING

One of the more important forward steps in recent social research is the investigation of qualitative coding as an *explicit* problem. The term "qualitative coding" is given to all the techniques for classifying reliably those social data on which very little order has been previously imposed by the researcher. Such a definition highlights the fact that coding is basically a matter of *classifying.* When the data have already been classified with respect to our theoretical needs, as was assumed in the previous section, our problems are mainly mechanical. However, when the data have not been structured in neat categories by carefully designed questions, the problems are more complex.

As has been noted several times, not only all social research but also all social interaction requires this kind of classification. We are constantly ordering the manifold social experiences in which we participate. When we investigate a past historical epoch, we must begin to organize data which were not recorded for our purposes. If we attempt to study the social patterns of a small community, or of a work unit, or of a primitive society, we must learn how to classify the relevant data.

Similarly, when schedules are tested, intensive interviews which are essentially unstructured are often used, and these answers are classified before a more structured set of questions is formulated. By this kind of approach, it becomes possible to use important sources of data beyond those already patterned by narrow sets of questions. Historical records, case records of various clinics or agencies, diaries and letters, propaganda and advertising, or nondirective interviews may thus become grist for the researcher's mill.

Most research based on such records has, however, been open to the criticism that the reader did not "see" the order created by the researcher. The conclusions were based on perhaps adequate data, but the reader did not agree that the data had been interpreted correctly. Yet each researcher must eventually meet the criticisms of his peers, and unless he sets forth a set of instructions for classifying the same data, no one is bound to accept the results. "Intuition" is useful in the search for creative ideas, but the *demonstration* of their truth cannot be founded upon any expectation that the intuition of others will come to the same conclusions. It is therefore necessary to develop an *explicit* set of instructions for handling data, thus forging a link between apparently unordered observations and adequately demonstrated generalizations.

Research experience in this area has not, however, been adequately organized and published as yet. It will therefore be useful to outline here some of the main steps which are ordinarily followed in qualitative coding, along with some of the problems which may arise.

1. CLARIFY WHAT IT IS THAT IS DESIRED FROM THE MATERIALS. Let us suppose that pilot interviews concerning residential mobility have used the question, "Would you mind telling me, in detail, just how you came to decide on moving from your last address?" If the interviews were careful, there should be several remarks from each respondent. Presumably, these answers represent many different *types* of decisions, or housing situations, or families, or even modes of participation in the community. That is, since the question is only partly structured, the answers may be ordered in many ways. How should they be classified?

The answer must be found in the purposes of the question. If the question was considered ambiguous by the respondents, the answers may be studied only to see whether it was understood accurately. However, if the question has been developed beyond this stage, from among the many ways of classifying the answers, the one which best suits the aims of the total research is chosen. If we are mainly interested in *who* makes such decisions, we will note which persons are mentioned. We may instead seek cues to the *process* of reaching the decision and look for statements about its phases. If we are seeking *motives,* comments which attempt to justify the move will be sought. If we are seeking *predispositions,* we will attempt to classify the answers according to the descriptions of the previous housing accommodations, or according to the housing values expressed.

2. STUDY THE COMPLETED SCHEDULES CAREFULLY. This is an obvious second step. It is sometimes a disillusioning process, since the answers may seem less rewarding on the schedules than they seemed while the interview was under way. However, it is these records on which the demonstration of the facts must be based. If there was more in the interview situation than appears on the schedule, it is wise to attempt to develop better ways of recording such data before proceeding with the research.

3. WORK OUT THE CLASSES AND THE INDICATORS OF THE CLASSES. After closely studying the data, it is likely that a rough idea of the classes will be developed. In some cases, of course, these classes have been in mind from the beginning. In others, the possible groupings will only gradually take shape.

However, it is at this point that a first approximation of the *indicators* to be used in coding should be made. Since the respondent has chosen his own words, they will not fall into neatly prearranged classes. Slang words may appear. Both simple and complex rationalizations may be recorded. It is necessary to work out both the *classes* and the cues, comments, or phrases which are accepted as *equivalent* indicators for those classes. For example, housing respondents may be grouped into (1) those who are anxious to move; (2) those who are looking for a definite type of housing; and (3) those who are confronted by a housing "opportunity." This, of course, is only one type of classification, which may be used to interpret other comments on the process of decision. However, the respondents may not have used such phrases. The solution exists, then, in developing *coding instructions* (for the researcher himself as well as for others) which list the *indicators* for each class. For example, one respondent may not use the term "anxious" but may state, "What a dump that was! But we had to find some place to live, so we put up with it for a year." Or, "My cousin picked it out while we were in Idaho. We were certainly sick when we saw it, but what could we do?" Both these comments indicate the respondents had a strong predisposition to move and were dissatisfied with their housing from the beginning.

On the other hand, a clear indicator for the "opportunity" group might be any comment which suggested general satisfaction with the housing, until a "bargain" was discovered accidentally. This type of comment would be the clearest case, and further details could be used to delimit the class in the case of respondents who were not very satisfied but who state that they would not have moved if the new housing had not been found unexpectedly.

The goal, therefore, is to work out a set of instructions which will enable others to classify these comments into the types which are relevant for the research. Since, ordinarily, exactly the same wording cannot be expected, the practice is to look for equivalent *meaning*. To do this, the phrases and cues which will unequivocally indicate the proper class are listed and checked by seeing whether several different coders using them will secure the same results.

In some cases, the indicators are known from the beginning. For example, one might be interested in discovering the extent to which nursery-age children use various ethnic stereotypes in their play language. The records might consist of observations made through a one-way visual screen, supplemented by wire recordings. In this case the coding problem would be simple, for the phrases used in ethnic stereotypes are fairly well known.

A somewhat more difficult problem occurs when a *continuum of intensity* is the object of the coding operation, such as the classification of answers from those expressing a strong degree of disapproval, to those expressing strong approval, of a particular object. As noted in the sections on scaling, this problem is sometimes met by the use of judges, who classify independently a series of responses along a continuum of approval and disapproval.

Even when judges are employed for a preliminary group of responses, the problem of which *indicators* are adequate for classification must still be faced, for we are dealing here with *unstructured* data. How many degrees can be used will depend on the detail to be found in the responses. "Strong disapproval" or "strong approval" may be fairly easy to classify. The less intense responses will sometimes be indicated by favorable or unfavorable comments which are accompanied by various qualifications: "They're all right, but some of them . . ."; or "I don't have much to do with them, but I never have any trouble. . . ."

4. FIT THE CLASSES TO THE DATA. Having worked out a preliminary set of indicators, the gradual process of fitting classes to the data must be begun. A first use of this procedure is to discover whether or not the instructions are clear. By working out the preliminary indicators and applying them, it is often possible to add new details to the instructions, as well as to understand more clearly the nature of the classes.

At this stage it is particularly important to pay attention to the *deviant cases*. No matter how carefully the first set of instructions have been worked out, some responses will occur which do not fit. The indicators may not apply, or they may seem fuzzy and vague. These cases should be studied in detail, for they may require a new set of indicators, or a new set of classes.

Further, it is likely that a small group of "unknown" or "unclassifiable" respondents will remain throughout the coding process. If this group is large, the instructions, classes, or data are poor. On the other hand, if this class is minute, it will not affect the conclusions to any important degree.

5. CODE ALL THE ANSWERS. Having gradually fitted the instructions to the classes needed, an attempt to code all the answers should then be made. Here, again, it is necessary to test the coding of one person or group against that of another. It is only by such fitting and testing that it is possible to check whether or not merely impressionistic judgments have served as the basis of interpretation.

Special problems. Before showing how qualitative coding is applied to communication content, such as newspapers, radio programs, speeches, etc., some common problems which occur often will be discussed.

A common result of the attempt to code is the discovery that one has failed to clarify the *concept itself* by which the comments are to be ordered. For example, a question about "belief in democracy" may elicit comments which are difficult to organize, precisely because the concept itself has not been properly defined. In some cases, the problem can be solved by a series of questions or by wording them more precisely. In other cases, it may be necessary to think through the problem once more. For example, if an attempt to order the comments in terms of "shop morale" confuses job satisfaction, acceptance of group standards in the shop, belief in job mobility, or other possible meanings of this term, only confusion in coding can result. In such a case, the focus of the research must be more clearly defined.

The failure to elicit the data desired is another type of problem which is sometimes met. Usually, this will be caused by faulty wording of the original question. Thus, one might use a direct question, such as "What made you commit this crime?" when the aim is to obtain information on the *process of deciding* to undertake a *particular* criminal act. Many respondents, however, will answer in terms of familial backgrounds, slum neighborhoods, gang behavior, etc. Then, of course, it is necessary to formulate the question differently.

Coding may result in an equally disheartening discovery—that most of the answers prove to be of one type. Thus, the earlier question about residential mobility, "Why did you decide to buy a house?" may elicit a set of answers which may all be classifiable as cost factors. Other questions indicate clearly, however, that cost cannot be the main factor. Respondents do not know much about comparative costs, and most of their comments suggest that familial discussions did not deal with costs to any great extent. We then see that "cost" is a kind of rationalization for other motivations, in terms which are acceptable in conversations with friends and outsiders. The answers to the formal question may have been so nearly identical merely because the basic question was not really being answered.

Such "one-class" tabulations often result when the question is answered by a cliché: "I decided to become an engineer because I've always been interested in it." "We were married because we were in love." "I never punish my children because they must be allowed to express their personalities." Reformulation of the question, or the use of probe questions, may be required. In any event, the coding process clearly points to a set of answers which seem to require further analysis before being accepted as satisfactory.

On the other hand, the attempt to find order in essentially unstructured data may lead to *further* insights, useful in understanding the tabulations already made. For example, in one unpublished study there were questions which dealt with ethnic opinions and friendship patterns. These answers were adequately coded in terms of attitudes toward various ethnic groups and also in terms of friendship patterns. However, this process suggested that the comments contained further data. When the respondents spoke of their friendships with members of other ethnic groups, they made "individual exemptions"; *i.e., their* friends were *different.* Their Irish friends were not pugnacious or of a low culture, or their Jewish friends were not commercial-minded (or radical). Consequently, a further coding operation was carried out, in terms of the individual exemptions used. Later cross tabulations demonstrated the fruitfulness of this operation, since a relationship existed between the degree of ethnic prejudice and the types of individual exemptions made.

Attempting to make procedures explicit is merely following one requirement of scientific work—that operations must be repeatable by others. If interview protocols are seen as "observations," it is clearly necessary that these observations actually be classifiable by others as they are by the researcher. When these observations are neatly structured, the problem of classification has been solved. However, many records such as case records in psychological clinics or social-work agencies, or observations of children, or protocols of free answers to open-ended questions, may be closer to the "real" attitudes of our sample. Qualitative coding is one set of techniques for locating and specifying the order which exists in such materials, and to which our theoretical questions point.

CONTENT ANALYSIS

We shall not at this point attempt to introduce the *substantive* field of communication analysis, which has had its main development within the past generation. All that we wish to do is point out that the data from all fields of communication may also be treated as sociological data. Whether we wish to understand the T'ang Dynasty in China or the Protestant Revolution, the propaganda of revolution or of reaction, the radio programs of the lower class housewife or of the highbrow, the available materials must be somehow coded. Whether these materials support or weaken a hypothesis in communication theory can only be determined by the usual criteria of good research design. All that is attempted here is to show how such materials can be exploited *as observations,* so that their underlying order, if any, is clearly demonstrated.

A simple case from the field of communication follows, in which such coding is indispensable for content analysis. While measuring the kinds of stimuli to which the newspaper public was exposed during a given period, a researcher may come to the conclusion that a particular newspaper was in sympathy with the Nazis at that time. A number of people may object to this statement, pointing out that the owner is a respectable man in the community, that he is "a 100 per cent American," that he does not belong to any political party, and so on.

If the researcher has been following a merely intuitive analysis, he would have to argue the point interminably. If he cites an example, they may cite a counter-example or an explanation. He may be charged with bias, with not understanding the situation, and so on. Many of our daily arguments with friends are similarly inconclusive.

However, a good sampling plan for the newspaper issues analyzed, an adequate logical structure uniting the data and the conclusions, and a systematic technique for coding the content of the newspaper make it possible to provide genuine evidence, such as Table 1 contains. With this table might go a comparison with other publications, such as the second tabulation here.

TABLE 1

Number of Statements in *The True American* Consistent with or Opposed to Nazi Propaganda (March 3, 1943 to December 31, 1943) *

	Consistent	Opposed
The U.S. is internally corrupt	301	34
The foreign policies of the U.S. are un-Christian	41	0
The President of the U.S. is objectionable	150	17
Great Britain is internally corrupt	34	0
The foreign policies of Great Britain are un-Christian	80	2
Nazi Germany is just and virtuous	56	2
Japan's foreign policies are morally justifiable	22	6
Japan is powerful	30	2
The U.S. is weak	345	41
Nazi Germany is strong	41	2
The U.S. and the world are menaced by:		
a. Communists	55	0
b. Jews	99	0
c. Meddlers in the affairs of other countries	45	0

* See Harold D. Lasswell, Nathan Leites, *et al., Language of Politics* (New York: Stewart, 1949), Chap. 9, "Detection: Propaganda Detection and the Courts," pp. 173–233, esp. the table on p. 187. Table above is adapted from Lasswell's presentation.

TABLE 2

Percentages of Articles and Editorials Consistent with Nazi Themes •

Magazine	Percentage
Time Magazine	0.11
True American	55.67
Liberty	1.12
Saturday Evening Post	0.54

• See similar tables in Harold D. Lasswell, Nathan Leites, *et al.*, *Language of Politics* (New York: Stewart, 1949), pp. 198–199. This table is an example of Lasswell's "consistency test"; the former, of his "parallel test."

Without going into the logic of the research design, which was developed during the war and used for actual court proof, we can see that such pinpoint use of data is at the polar extreme from an intuitive approach. It will be noted that the researchers have actually counted the presence or absence of Nazi themes within a publication. Whether the focus is on editorials alone, or on the entire publication, these examples show a translation from the qualitative complexity of the total content, to the quantitative simplicity of counting Nazi themes. Since these themes are already available, the coding problem would not be difficult. For high reliability, coding instructions would simply list the major themes to be observed, and coding the publication over a selected period would yield an accurate count, which could be verified by other researchers. Of course, not all the material from this publication is utilized. Many other analyses could be made, with relevance to other problems, and "theme" analysis is only one possible unit of investigation.

Let us now select another example, dealing with the status of various ethnic groups in the United States, as depicted in popular fiction. A great number of interesting critical essays could be written on the subject, depending on the writer's literary knowledge. These might be suggestive and stimulating, but most of them are open to the criticism that the reader "knows of exceptions," or he "does not see" how the writer arrived at such conclusions. By developing an adequate sampling plan and a set of clear coding operations, it is possible to demonstrate a number of facts about popular fiction which do not depend on the researcher's personality or insight alone but which can be tested by any trained social scientist.

One study of this sort focused on the treatment of ethnic groups in popular short stories.[2] The method of classifying the data was fairly simple. Each story has a number of characters, and each of these characters is described more or less fully. The student can extract this list of characters with their descriptions from each story.

---

[2] Bernard Berelson and Patricia Salter, "Majority and Minority Americans: An Analysis of Magazine Fiction," *Public Opinion Quarterly,* Vol. X (1946), pp. 168–190.

Any such set of characters can then be analyzed, in order to discover whether there is any identification of class, race, ethnic origin, religion, occupation, the goals which they seek, and so on. The process of developing explicit coding operations, as noted previously, depends on the purpose of the analysis and on the difficulty of securing adequate indicators for the desired classes. For the case in question, let us take a few examples of hypothetical characters as met in popular fiction. A simple decision might be reached on the following individual, who is presented with the description used in the story about him:

Character	*How Identified*	*Treatment in the Story*
John Adams	White (story illustration) Descendant of New England family Lawyer High income (has custom-built convertible, seashore summer cottage) Episcopal (has conversation with minister regarding early church experience)	Hero of the story Idealist (is trying to force realtors to accept a city plan which will beautify the community, but reduce their income)

Similarly, the following character seems adequately identified to allow some judgment concerning his treatment in the story:

Character	*How Identified*	*Treatment in the Story*
Johnny Turridu	White Came to this country from Italy at the age of three (father speaks broken English) High income (spends vacations at expensive resorts, owns buildings) Catholic	Minor role. He is defeated by hero Makes money from numbers rackets Rejects his old priest Cynical materialist (expresses interest only in his success; bribes city officials)

Any such sketches would, of course, be only the first steps toward more precise directions for identifying the ethnic backgrounds and the roles which the characters play. There would be fewer data for minor characters, and instructions must be flexible enough to allow identification by using any one of the many indicators which might be used. For example: (1) the ethnic membership might be specifically noted; (2) obvious ethnic names might be used; (3) some supposedly ethnic phrases may occur; (4) membership in ethnic organizations may be cited; or (5) the home neighborhood may be described closely enough to allow identification. A similar set of possibilities exists for "treatment in the story," although this is somewhat more difficult, involving as it does a set of cues to locate and specify to what extent there is "discrimination" against the ethnic group member. So long, however, as the category, "unknown" is not a large one, the basic conclusions will not be affected.

Note, further, that for many minority groups there may be no representatives at all. This is not surprising, since popular fiction must rely upon commonly recognized ethnic images. Consequently, in this case the researchers used these categories: "The Americans," "The Anglo-Saxon and Nordic minorities and foreigners," and "The Others" (Jews, Italians, Negroes, etc.). The treatment of these groups in the fiction sample was then tabulated as shown in the accompanying table. "The rule seems to be that the character receives better treatment the closer he is to the norm of *The American, i.e.,* white, Protestant, English-speaking, Anglo-Saxon" (page 186).

TABLE 3

A Content Analysis of Ethnic Groups in Current American Literature

	The Americans, Per Cent	The Anglo-Saxons and Nordics, Per Cent	The Others, Per Cent
Major characters	52	38	30
Approved characters	80	78	62
Highest socioeconomic status	39	24	16
Top occupations	59	29	20
"Heart" goals (*i.e.,* idealist, love, family, as against money, power, etc.)	69	61	49

Thus, by working out fairly simple rules for identification and treatment, it was possible to determine that within popular short stories there is a demonstrable "discrimination" against the various minority groups, especially those which are at present disadvantaged in the society. By using fairly large categories, the problem of decision was reduced considerably.

The data contained in the short stories were not originally presented or organized for the purposes of research, but by formulating instructions for locating and identifying the characters and the themes which are used to describe them, the coders could *reliably* classify the data as presented. The further analysis of the meaning of these data will always depend on the goals of the research, but the data are no longer guesses or hunches: they are empirically verifiable by other observers if the same operations are followed.

It is also clear that there are many *units* of the communication content which can be used for sampling and analysis, from small units such as words or symbols, to journal issues or volumes. Similarly, there are many ways of organizing the materials: by symbol count, by item analysis, by thematic analysis, in terms of the *structure* of the ideas and facts presented, or even (in propaganda materials) in terms of the campaign. Clearly, it is equally possible, in the case of content addressed to an audience, to analyze these spoken or written materials in terms of the *response*. However, for such further treatments of content analysis as a substantive field, the student should read the literature cited at the end of this chapter.

THE CASE STUDY

The case-study approach may be introduced at this point, since traditionally it has been thought of as a kind of qualitative analysis. More important for our purposes, however, is the fact that it is through the developments of qualitative techniques, such as qualitative coding, that the case study can be most efficiently used in social research.

References in the sociological literature to "the case-study method" are decreasing. In part, this would seem to be one result of the false distinction between "the statistical approach," and "the nonstatistical approach," as a consequence of which the case study is sometimes identified with the use of less reliable research techniques. It is often thought of as a kind of intuitive approach, derived from much participant observation and using all sorts of personal documents such as diaries, letters, autobiographies, etc., without adequate sampling design or checks on bias or distortions resulting from personal views of social reality.

However, such an identification fails to distinguish the case-study *approach*, a method of looking at social reality, from the specific research *techniques* which accompanied it in the research of a generation ago. It is true that much of case-study research at that time failed to follow good sampling design and often came to conclusions with no explicit description of the operations which led to those conclusions. However, this was true for most social research, and in any event there is no intrinsic connection between the case study and unsophisticated research techniques.

Similarly, the case-study approach has been praised as a set of useful research techniques, such as the use of personal documents, probing many facets of the respondent's life, adjusting the wording of questions to fit the peculiar experience of the respondent, sharing experiences with the individuals being studied, collecting life histories, gathering data over the lifetime of the individual, etc. Such praise is equally unwarranted, since these techniques may be used in almost any type of social research. The use of qualitative coding permits the use of personal documents in a systematic way. The interview guide is one form of the questionnaire which allows flexible and alternative wordings of questions, when that is needed. Participant observation, while it has not been systematically studied in recent years, is used when possible in the early stages of formulating the research problem. Life histories may also be used in the preliminary stages of research, or as the basic body of data, and may be employed in conjunction with the techniques of qualitative coding for many types of social research. In short, modern social research has attempted gradually to systematize and make precise the various special "qualitative" techniques which were once thought to be the exclusive characteristics of the case-study approach.

The case study, then, is not a specific technique. It is a way of organizing social data so as to preserve the *unitary character of the social object being studied.* Expressed somewhat differently, it is an approach which views any social unit as a whole. Almost always, this means of approach includes the *development* of that unit, which may be a person, a family or other social group, a set of relationships or processes (such as family crises, adjustment to disease, friendship formation, ethnic invasion of a neighborhood, etc.), or even an entire culture.

Before pointing out some of the difficulties and advantages of this approach, let us clarify it somewhat by contrasting this holistic view with that of much modern research. A simple case taken from public opinion polling yields an extreme comparison. Suppose the respondents are asked a series of questions about their attitudes toward war. For further analysis, a series of cross-tabulation questions, such as age, sex, economic ranking, urban-rural residence, and veteran status, may also be asked. The answers to each question are classified, in order to cross-tabulate by strata. Thus the answers to the question on economic ranking are put into the classes A, B, C, and D. Perhaps three to five age groupings and two sex groupings (male and female) may also be used. Having thus classified and coded the data, all these can be transferred to the punch cards.

However, from this point onward the individual unit, the *person,* disappears from the analysis, which instead merely compares *traits.* Thus, for example, cross tabulations may reveal whether the old object to war more vigorously than do the young, or whether women disapprove of war more vigorously than do men, or perhaps how the various economic rankings compare in their attitudes toward war.

The answer to each of these questions is obtained by sorting out all the cards into the first group of traits (old as against young), then tabulating under each group the responses to questions about war. For example, the cards could be sorted into those with a punch meaning "65 and over," one which means "under 20," and another group which includes all the rest (or any other age classification desired). Then, for each such group obtained in this fashion, the responses to the questions concerning war are tabulated.

Note, however, that it is only the *traits* which are immediately involved in the cross tabulation. All other characteristics on each card are ignored. What is being emphasized here is not the *narrowness of the questions,* for even the most elaborate research represents a relatively thin slice of any respondent's life. Rather, it is important to see that most of our analyses are cross-tabulating the distribution of one *trait* with the distribution of another. This is equally the case for elaborate cross tabulations, as it is for correlations. The individual representing the "wholeness" of those traits does not figure in these analyses.

Such a characterization is, of course, essentially negative and seems almost a polemic cliché. In order, then, to specify the case-study approach as an alternative mode of handling respondents, groups, or social relationships, let us attempt to show wherein the differences lie. In a more practical form, how *can* the wholeness of cases be preserved? We shall attempt to outline the answer briefly under four rather complex rubrics: (1) breadth of data; (2) levels of data; (3) formation of indexes and types; and (4) interaction in a time dimension.

A few preliminary remarks are in order, however. First, the *wholeness* of any object, whether physical, biological, or social, is an intellectual construct. Concretely, there are no limits which define any process or object. Every variable ultimately links with any other. As theoretical biology has pointed out, from some points of view even the living animal is a construct, and the point at which the animal stops and "the environment" begins is arbitrarily defined. Similarly, "the individual," or "the formation of a gang" must be such a construct, defined in terms that are most fruitful for the research problem.

From this point we see, further, that not only is it difficult to set the *limits* of any social object, but it is difficult to state at what point it is profitable to stop gathering data about the object so delimited. As has been noted before, an infinite number of observations may be made about even the dullest or most unimportant person. Consequently, neither the case study nor any other approach can be characterized as the analysis of the individual in all his uniqueness. The point at which data are adequate must be determined by the research problem itself. There is no inherent or intrinsic limit. One implication of this must be added: when we speak of the "social object as a whole" we cannot mean the "social object as a unique unit." It is precisely the unique elements in any phenomenon which are eliminated by scientific abstraction. To the extent that characteristics are really unique, it is impossible to account for them in terms of scientific generalizations. Consequently, the case-study method cannot be thought of as capturing the unique, but as attempting to keep together, as a unit, those characteristics which are relevant to the scientific problem being investigated.

Breadth of data. Clearly, one of the ways by which the unit is studied as a totality is the collection of a broad array of data about it. This is true whether the subject is a set of social relationships, a person, or a group. The "person" who is recorded on a polling schedule is not only dissolved by becoming a set of traits individually tabulated; he almost fails to exist from the beginning, by virtue of the narrow range of the data concerning him. Although mere quantity of data is not sufficient, since the collection must be guided by the research problem, there is a greater opportunity to grasp the pattern of the individual's life if a substantial body of data concerning many facets of that life is available.

Levels of data. The case-study method is further distinguished by the use of data from other abstract levels than the purely sociological. Although many sociologists do incorporate economic, political, psychodynamic, and even biological data into their analyses, social research has in the main been moving in the direction of a clearer definition of the sociological, together with better techniques for abstracting it from the complex of other relationships. In the case-study method, such data from other levels are deliberately incorporated into the records. Although it may be pointed out that such data should be viewed in terms of their sociological meaning, so that a biological defect, for example, is seen as it affects social relationships, it is clear that recourse to other levels does give added dimensions to the individual being studied. When we see the individual in his total network of relationships, it is more difficult to lose sight of him as a unit.

Formation of indices and types. However, the most important technique in preserving the wholeness of the social unit is the development of typologies and indexes, so that the various traits are actually used in characterizing the unit. Note that this technique is not confined to the case-study method and is used in any qualitative analysis. Furthermore, elaborate cross tabulations attempt the same goal. Thus, cross tabulations in terms of sex, age, and socioeconomic ranking serve to characterize those who are (for example) male, *and* aged 20 to 29, *and* poor.

The same operations may be performed in a case study, and the problem may be somewhat simplified by qualitative coding beforehand. Because of the breadth and added dimensions of the data, moreover, the choice of *which* type or class the case falls into will be simplified. In addition, by the time the types begin to emerge clearly, enough of the process of interaction between various facets of the individual's life will be known so that this choice may be made with greater certainty. Finally, cases will be concerned with developing "natural types," that is, types which are already discerned in the culture (such as "lovers' quarrels," "life of the party," "village philosopher," or the religious revival meeting), and the case-study approach can ordinarily single out such types at an early stage.

Interaction in a time dimension. To the breadth and additional levels of the data gathered in the case-study method must be added the emphasis upon *process and time.* Any cross tabulation is essentially a static analysis, although the relative numbers of cases found in various subcells may be used to *test* hypotheses of interaction. For example, according to some theories we would expect to find a causal relationship between class position and continuation in school. It is possible, then, to cross-tabulate these characteristics (possibly holding IQ constant) and thus check the hypothesis.

Nevertheless, we do not observe "process" in such tabulations. We are limited to the traits on the punch card, which have not recorded any of their relationships to one another. In the case study, the attempt to hold these characteristics together in both the data-gathering and the data-analysis phases emphasizes the *changes* in time, as well as the *processes* by which those changes took place. The period of time may be short or long. An individual's life from childhood until the time of a study, or the formation and reformation of cliques in Congress over a very short period of months may be studied equally well. In either case, however, the concern is with recording the relevant characteristics *as they appear* in interaction, not merely recording them at two separate instants in time for a before-and-after comparison. The emphasis on the interaction process makes the maintenance of the unitary character of the social object somewhat simpler.

Problems of the case-study technique. Most of the difficulties in the use of this method can be reduced to one, although a more complex classification is possible. Interestingly enough, the basic danger in its use is the *response of the researcher.* The researcher comes to feel a false sense of certainty about his own conclusions. The danger, then, does not lie in any technical weakness of this approach to social processes or individuals as wholes.

The student may feel that this is a constant danger in all research. Yet it should be remembered that in most research there are constant reminders that a very narrow range of experience is represented by the data. However long our interviews, gaps are all too obvious, and there is no way to fill them. However, each *case* which is developed as a unit takes on complete dimensions in the mind of the researcher. He comes to feel certain that he could answer many more questions about his case than can be answered from his file records. The case has a definite form and pattern, and as the researcher probes more deeply into the process or person being studied, he finds there are few surprises left. It is quite comparable to our feeling of certainty about our close friends, or the neighborhood in which we have lived for years, or our families. There is, in short, an emotional feeling of certainty which is much stronger than in the case of other types of research. This is particularly true as compared with much survey work, in which the analyst has only the completed schedules before him and knows he cannot capture the varied experiences of the many interviewers who carried out those interviews.

Yet the student can test the accuracy of this knowledge very easily, by submitting to questioning about his close friends or his own family, the "cases" he knows best. Most of us find that our feelings of certainty deceive us. We know far fewer data about these cases than we believe. Important facts are forgotten. Other facts are distorted, or surprising facets of these lives are unknown to us. For example, children usually believe that they "understand" their parents, but are often oblivious to the many deep conflicts between them. Brothers and sisters are often unaware of important experiences happening to each other outside the family circle. What most of us feel is a kind of at-homeness with these close acquaintances, so that we fill in or ignore those facts or dimensions of the person which are really unknown to us.

This danger, then, is one which the observer *himself* creates. The *consequences* of this feeling of certainty are many, but most of them can be grouped under one main heading: *a temptation to ignore basic principles of research design.* Since the researcher feels so very certain about the area of experience he is investigating, he feels no need to check the over-all design of proof. For example, after the student has collected, say, 200 cases of juvenile delinquency from social-work records, supplemented by interviewing and other sources, it is difficult for him to feel that he does not have an adequate sample. The range of delinquency experience in his cases is so wide, the types of people so varied, the depth of detail so vivid, that the researcher ordinarily has a strong conviction that his selection is "representative." He was not, usually, following any known pattern of sampling, so that he feels certain there is no bias. It must be repeated at this point that we are not separating one type of researcher from any other type. *Any* investigator who absorbs the facts from a large number of cases will begin to feel that he really has a satisfactory sample, no matter how much knowledge he has about sampling design.

The result is, naturally, a strong temptation to extrapolate unwarrantedly. Perhaps an equally important consequence of this feeling is the failure to make explicit just what are the generalizations underlying the analysis of the cases. One boy's refusal to give up his pilfering may be explained by his desire for economic gain. Yet, in a similar case, and without explaining the *bases* for the different explanation, another boy's refusal to give up thievery may be explained by his desire to humble his parents, a psychodynamic explanation. Although both explanations may be correct, we are likely to forget that any such analysis requires a *system* of explanation. It is not scientific to apply various explanations which by common sense or intuition seem to fit particular cases, on a purely *ad hoc* basis. Some set of generalizations is being used, and it is necessary to make explicit just why one or the other may or may not be applied. However, when the investigator feels so much "at home" with each case, he may be tempted to "feel" the right explanation—even though a second reader of our cases might come to an entirely different explanation. This is what is known as *ad hoc* theorizing. The previous reference suggests a further danger resulting from this feeling of certainty: *the failure to test the reliability* of the data recorded, the classifications used, or the analysis of the data. As the researcher is likely to point out, no one else knows the data nearly so well as he, so that no one else could check his work adequately. Furthermore, case gathering is a time-consuming activity, and it is difficult to find others who are willing to study the cases with such completeness.

Avoiding these problems. Nevertheless, these difficulties can be avoided by the student who is willing to follow good research planning. Being warned of the dangers which result from such a feeling of certainty, he will develop a research design which takes account of these dangers. He cannot avoid the feeling of certainty, but he can use an adequate sampling pattern. Knowing that his sampling is good, he has a rational basis for making estimates about the universe from which it was drawn. Instead of relying on intuition for his conclusions, he will make certain that whenever he analyzes a given factor, there are actual data for all case records concerning that factor. Further, he will develop his theoretical framework from the beginning of the research, so that he avoids speculations made on the spot to fit the peculiarities of each case.

Furthermore, he will attempt, as far as possible, to utilize the technique of qualitative coding for individual factors and traits which are amenable to such classifications. If he is to use categories like "selfish," or "adjusted," or "anomic," he will develop a set of instructions for deciding whether a given case falls into the category. If other investigators cannot use such instructions, they are likely to contain many implicit judgments, intuitions, common-sense guesses, and the like. The good student will recognize that a description of research operations must be so written that other scientists can repeat them. If not, either the description or the operations are unclear. The student *must* make this test, if he is not to remain open to the charge from others that they "simply can't see" what he sees. It is usually good procedure to carry out this stage of analysis with the aid of a number of collaborators or assistants who can act as judges of the reliability of even simple classifications.

One may almost claim that it is at this point—that of developing categories, of defining and delimiting types of behavior already recorded—that the case study requires the closest criticism. For it is at this stage that the systematization of much of the mass of data begins, that is, condensation, excision, reinterpretation, etc. The final report cannot simply repeat the recorded observations. The steps in this transformation, then, must be carefully marked, if the researcher expects to have his conclusions accepted.

It is worth recording that this problem is not peculiar to sociology. When the historian begins to reduce the mass of individual observations from original sources, he faces the problem of showing his operations to fellow historians, so that they can follow them exactly, should they wish to do so. Much of the objection of social scientists to the published analyses of psychiatrists and psychoanalysts is simply that, whatever the details of the case, there is no clearly outlined set of operations by which the next scientist can use the same data to come to the same conclusions. The economist who studies the marketing process among tobacco farmers may find similarly that each case is so very different that colleagues are suspicious of conclusions until the operations of the analysis are made clear. As noted above, the best way of clarifying these steps is to develop definite coding instructions for most of the important qualitative items, with definite tests for coding reliability.

Costs in time and money. It is clear that this approach is a costly one in time and money. Each case becomes a research in itself, and the collection of even 100 adequately documented cases may easily consume 2 years of the student's time. Since the data are not usually standardized, the system of keeping records and of developing techniques for standardizing the observations must be rather elaborate. Over such a lengthy period, the loss of potential cases from the sample drawn may be very great, and the resultant costs of tracing cases similarly great.

These costs need not deter the researcher from this technique. They simply remind him of a cardinal fact, that *all* research is expensive, and there must be an early judgment as to *which* kind of study will be most adequate for the kinds of facts needed. For some types of analysis, the use of case studies will be fruitful. For others, a simple polling study or mailed questionnaire may be adequate. When detailed *processes* of social interaction are being analyzed, some preliminary recourse to case study is needed, even if there is not complete reliance upon this method. One type of simple, less expensive, substitute is a panel sample, which is reinterviewed over a period of months.

Advantages of case study. Nevertheless, as was noted previously, most good research has some recourse to case analysis, even if it does not appear in the final publication of results. Its fruitfulness is indicated by the preceding discussion of its characteristics. Before a final questionnaire is developed, it is likely that the careful researcher will attempt the analysis of many cases which center on his major problem. He cannot rely on the reports from a few trial interviews. He will want to go over the field of interaction with a number of respondents, in great detail, to find out just how their life pattern is being affected by the item being studied. For example, if the aim is to develop a good schedule for the analysis of stratification, simply trying out a number of schedule drafts will usually not be adequate. The student must attempt to develop a body of much more intensive, detailed, and subtle observations about many individuals, learning about their responses to the phenomena of class barriers, mobility, class perspectives, etc. In addition, it is here that the added data allow the study of the deviant cases, that is, those which seem to negate the theory. Such analysis may lead to considerable clarification of the theory itself. Such observations may or may not be formalized into definite case histories. Such a practice is recommended, since at this exploratory stage the student is likely to see *selectively*, that is, see those items which seem outstanding to him, while neglecting perhaps equally important data. Only by *recording* them for analysis and then analyzing them, prior to the next stage of formulating the final version of questionnaire or schedule, can he make certain that all the pertinent data are being incorporated into the research instrument.

Of course, the researcher may be interested in certain social processes, such as the courtship process, clique formation, postdivorce adjustment, or boys' gangs, and in such a case the unit selected need not be the life history of the individual person. It will instead be the *process* selected for study. Here again the preliminary study of the processes is greatly facilitated by case documents, incorporating the pertinent data which will guide later stages of the research.

Whether the student approaches his data through the framework of the case study at *later* stages of the investigation will depend on many factors. As was noted above, the time and money at his disposal may lead to a decision to study only a limited number of factors. In this case, he may well utilize a questionnaire or some form of single interview as his main technique. If he wishes to study a much wider range of data, he may be willing to limit the number of cases. He will ordinarily not be able to generalize safely from a small number of cases, but it is often true that the depth of insight afforded by case study will yield fruitful hypotheses for a later, full-scale study.

The most important theoretical advantage of the case study has not been fully exploited as yet. As noted earlier, most research has become "trait" research, that is, cross tabulations between *traits* of individuals. Even when these individuals are being studied in groups, very few individuals or groups as units are being analyzed.[8] Even though many of the apparently group traits can be approached through combining individual data, they may also be approached through an organizing framework emphasizing the characteristics of the group, or process, or social structure as a whole. The case study, attempting to organize data around the unit of growth, or group structure, or individual life pattern, does force the researcher to think in these terms rather than fall back on trait analysis alone.

---

[8] For a recent discussion, see Patricia L. Kendall and Paul F. Lazarsfeld, "Problems of Survey Analysis," *Continuities in Social Research*, Robert K. Merton and Paul F. Lazarsfeld, eds. (Glencoe, Ill.: Free Press, 1950), pp. 187–196.

Emphasis should also be placed on the *wider range of personal experience* which the use of the case study gives to the student. This advantage has been implicit in the previous discussion, but it deserves special comment at this point. Precisely because of the narrowness of most survey work, the researcher actually derives most of his wider range of experience in such investigations at the stage of *analysis,* when the meaning of the questions is probed more deeply. This latter stage is most fruitful, however, if there has been a *prior* period of absorbing the varied experience of others. The case study is particularly useful because of its attempt to find the meaning of the recorded data within the life of the individual, and only later in terms of *classes* of individuals. Often, too, these experiences yield new insights because of their very differences from the ordinary experiences of the researcher; he may never have been divorced, been a safecracker, a pickpocket, a member of a boys' gang, a member of an ethnic group becoming Americanized, etc. In a sense, then, he is able thereby to live many quite different lives by sharing these experiences. Such materials are not only useful within the confines of his particular research, but may become raw materials for further reflection and research. Since the researcher, whether academic, industrial, or governmental, is under pressure to lead his life entirely within his own middle-class group, attention to these further dimensions of social life prevents his sociological thinking from becoming increasingly narrow.

It is seen, then, that although it is not possible to identify the case-study "method" as a particular *technique* for eliciting data, it is a *mode of organizing data* in terms of some chosen unit, such as the individual life history, the history of a group, or some delimited social process. In order to obtain such holistic data, one may use all the techniques which any other mode of organization uses: intensive interviews, questionnaires, self-histories, documents, case reports by others, letters, etc. Maintenance of the unitary character of the case is aided by the breadth and added levels of data gathered, the use of indexes and typologies, and the emphasis on interaction in a time dimension. There is, then, some attempt to make of each case a research in itself. This is a time-consuming process, and the collection of a large number of cases may be unnecessary, if the research goals are limited. The absorption in the detailed material of social relationships and interaction gives the researcher a wider range and a greater depth of experience, which may be of great use in interpreting the data and in further research. However, by such close contact with the individuals or groups studied, it is likely that the observer will have a strong effect on the data themselves. At least, it may be claimed that this danger is perhaps greater in this approach. The absorption in the data also creates such a feeling of certainty about his knowledge, on the part of the researcher, that he is often tempted to be less cautious in following the basic rules of research design. Nevertheless, for preliminary research in any field, most investigators will use some form of the case study; and for the purpose of *group or process analysis,* as against the analysis of individual traits alone, it is a highly fruitful approach, as yet insufficiently exploited by those who are currently doing research into research techniques.

William J. Goode and Paul K. Hatt

Thus far the emphasis has been on the necessity of carefully *planning* the collection of data so that significant questions can be answered and on the problems of *collecting* data so that the answers will be reliable and valid. It is quite obvious, however, that performing these two steps alone will not provide an answer to the original hypothesis. Rather, they will provide a means for answering it. This chapter aims at clarifying some of the ways in which such data can be so ordered as to provide the desired answers.

Under ideal conditions of precision and simplicity this presents very few problems since the statement of the hypothesis and the elaboration of the experimental design will automatically provide for the analysis of the data. Consider, for example, a hypothesis stated as follows: If lights are selected as visual stimuli in such a way that each is subliminal (not perceivable) to each eye taken separately, then if they are applied simultaneously (but still separately) they will be perceived by the subject. Such a hypothesis is so stated that either there will be perception or there will not. An analysis of the data would constitute no problem at all. If, on the other hand, the responses were such as to show not an absolute presence or absence but rather proportions or degrees of presence, the problem of analysis would become more complex. If, in addition, the proportions of those perceiving the light appears to be different among different types of subjects, *e.g.*, men and women, or older and younger persons, the problem becomes rapidly still more complex.

In other words, the problems raised by the analysis of data are directly related to the complexity of the hypothesis or hypotheses. Suppose, for example, a study is carried on to test the hypothesis that family size is directly related to size of the home in which the family resides, with the results shown in the table which follows. Such findings would force a rejection of the hypothesis *as stated* and should lead to its reformulation in more specific, and hence more complex, terms. Thus, it might be restated as follows: The size of the home is positively correlated with family size, *when the opportunities for choice are equal.* This, of course,

FAMILY SIZE	PERCENTAGE OF FAMILIES RESIDING IN:		
	Small Houses	Medium Houses	Large Houses
Large families	25.0	55.0	20.0
Medium families	30.0	60.0	10.0
Small families	30.0	55.0	15.0

requires a clear definition of what factors influence freedom of choice. Clearly, financial ability to buy or rent at various levels is such a factor, since size of home is correlated with its costliness. A retabulation, therefore, such as the following one might then appear. This type of finding would tend roughly to support the hypothesis as restated. If similar tabulations were carried out for "medium" and "small" families, and coincided with the above findings, then the hypothesis would be affirmed even more strongly.

ECONOMIC LEVEL OF FAMILY	PERCENTAGE OF LARGE FAMILIES RESIDING IN:		
	Small Houses	Medium Houses	Large Houses
High	5.0	35.0	60.0
Medium	35.0	65.0	10.0
Low	40.0	55.0	5.0

However, economic ability is not the only factor which interferes with the choice of a home. In some areas being a member of a "non-Caucasian" race will inhibit choice through restrictive covenants. Therefore, it might be necessary to retabulate the above in terms of race. Thus such an analysis would lead to two racial divisions (white and nonwhite) within three economic levels (high, medium, and low) for three sizes of family (large, medium, and small). This would yield 2 × 3 × 3, or 18 such tables. As will be seen later, it is not always necessary to perform the task in so cumbersome a fashion, but the logical problem must nonetheless be dealt with.

If all the relevant factors are known in advance, then there is no serious problem in analysis, for as pointed out in the earlier chapters, the experimental design would have gathered the data in only the crucial instances. There are two aspects of scientific research, however, which make the occurrence of so ideal a situation relatively infrequent. The first of these is the appearance of an anomalous empirical regularity or the absence of an expected regularity. Such instances require analysis not anticipated by the original design. Similar to this is the use of data, gathered for one purpose, for quite another problem. Both these represent what is called *secondary analyses.* From a very "pure" experimental point of view, such analyses are considered to yield answers which are "plausible" but not capable of being stated in the customary "probability" terms of science.

For example, in the case of the housing study referred to above, if race and economic status had been considered in the original design, the sample employed would have been so constructed as to contain an *adequate* and *representative* selection from all the various subtypes. That is, a satisfactory sample of each family size within each racial group within each economic level would have been secured. If this were *not* done because the later analysis was the result of an afterthought, it becomes difficult to assess the value of the old sample for the new purpose.

However, secondary analysis is not only common but necessary, and it is certainly of great value, even if it produces no more than plausible statements to serve as hypotheses for subsequent verification in more stringent terms.[1]

In a sense, then, problems of data analysis involve all the questions raised in the chapter on research design, for secondary analyses do involve the designing and redesigning of substitutes for the controlled experiment. This indicates that there are two types of questions about analysis which may be raised. One of them relates to the techniques of *representing the data,* and the other to the methods of *logically ordering* them so that questions can be raised and answered.

The first of these general problems cannot be treated exhaustively in a book such as this and are, in any case, part of the usual content of statistics courses. A few basic comments concerning the methods of representing the data will be made, but the major object of attention will be the logical processes involved in secondary analysis.

STATISTICAL REPRESENTATION

The frequency distribution. The simplest form of representing research findings is the frequency distribution or tabulation. All that is meant by this is the presentation in one column of different qualities of an attribute, or different values of a variable, together with entries in another column showing the frequency of the occurrence of each of the classes. The only problems connected with the preparation of a useful frequency distribution, or simple table as it may also be called, is to use common sense with respect to three things.

---

[1] A detailed discussion of this point can be found in *Continuities in Social Research,* Robert K. Merton and Paul F. Lazarsfeld, eds. (Glencoe, Ill.: Free Press, 1950), *passim,* esp. pp. 133–143 and pp. 197–211.

First, the units entered in the left-hand column describing the qualities or values must be mutually exclusive, as well as inclusive of the vast majority of observations which will be made. Overlapping values or attributes can only lead to confusion. For example, in a study of Puerto Rico one of the authors found that besides Catholic, Protestant, and "no religion" categories, there were some who professed Spiritualism. This would have made no problem except for the fact that it appeared possible to combine Spiritualism with either Protestantism or Catholicism. Therefore, instead of just adding a category, Spiritualist, it was necessary to add also Catholic *and* Spiritualist as well as Protestant *and* Spiritualist.

Second, the tabulation, to be of the most utility, must have *internal* logic and order. It seems quite obvious that, if one were tabulating such a variable as the height of men or the size of cities, he would tabulate in order either from the tallest or largest to the shortest or smallest, or vice versa. However, when tabulating qualities, where the order may not be so obvious, the need for a logical treatment is equally great. For example, one survey asked women for their reasons for buying a certain face cream.[2] One tabulation of the results is shown in the accompanying table.

Reasons for Buying Face Cream *

Reasons	*Percentage of Respondents*
Recommendation	28.0
Beneficial to skin	21.0
Heard it advertised over the radio	18.0
Saw it on the counter	15.0
Reasonably priced	10.0
Scent appealed	8.0
Because of special skin conditions	7.0
Total	107.0

* This is not strictly speaking a frequency, but a percentage distribution. This problem will be taken up later.

First of all, since the total reaches 107 per cent, it is clear that this table violates the first requirement, namely, that the categories be mutually exclusive. Ignoring this fact, however, it is clear that there is no sensible order to the types of reasons given. How is it possible to improve the utility of the tabulation for analytical purposes? Since there is no continuum such as that for height or size, the procedure is to search for logical groupings of the responses. It is possible to discover three such larger categories, and the use of the table increases greatly if it is modified in this way.

---

[2] Adapted from Hans Zeisel, *Say It with Figures* (New York: Harper, 1947), 2d ed., pp. 7–8.

Reasons for Buying Face Cream

Reasons		Percentage of Respondents
Pertaining to respondent:		28.0
Beneficial to skin	21.0	
Special skin condition	7.0	
Pertaining to product:		18.0
Reasonably priced	10.0	
Scent appealed	8.0	
Pertaining to way heard of product:		61.0
Recommendation	28.0	
Heard radio advertising	18.0	
Saw it on counter	15.0	

Not all tabulations will have a logical structure, but in most cases it is possible to find a logical order to use as an analytical principle, even in simple tabulation.

Third, when the left-hand column of a tabulation is a quantitative variable such as size of city or monthly rental value, the class intervals must be carefully and reasonably chosen. Schmid suggests three criteria for this: [8]

> "1. Ordinarily there should not be less than 8 or 10 and not more than 18 or 20 class-intervals, depending on the nature of the data and on the number of cases being studied. In order to obtain a clear understanding of the original data the individual items are frequently arranged in either ascending or descending order of magnitude. Such a classification is known as an *array.* After noting the highest and lowest values as well as the characteristic features of the data, the number of intervals can be determined more easily.
>
> "2. Every effort should be made to have intervals of uniform size. The intervals should not be so small as to lose the advantages of summarization or so large as to conceal the more important characteristics of the distribution. Moreover, if the class-intervals are too small, vacant or blank intervals might occur. If comparisons are to be made between similar data, it is advisable to select class-intervals of the same size for all the distributions. Whenever possible the class-intervals should represent common and convenient numerical divisions such as 5 or 10, rather than odd divisions such as 3 or 7.
>
> "3. After the size of the class-intervals has been determined, it is important that they be clearly designated in the frequency table. Each interval must have definite lower and upper limits, and must be expressed in such a way as to obviate any possibility of misinterpretation or confusion."

These seem like simple principles, and indeed they are. In practice, however, they are somewhat difficult, and many analyses of data have become difficult to perform or understand because of ignoring them.

---

[8] Calvin F. Schmid, "Basic Statistical Concepts and Techniques," in Pauline V. Young, *Scientific Social Surveys and Research* (New York: Prentice-Hall, 1949), pp. 291–292.

Summarizing the frequency distribution. There are two general ways of stating a frequency distribution in simple ways. These are used singly and together, depending upon the problem at hand. One of them is to compute a value which represents the *central tendency* of the distributions. Such measures are called averages and include among others the common average, technically known as the *mean;* the *median,* a value such that half the entries in a frequency table fall below and half above it; and the *mode,* or the value represented by the greatest frequency.

The other general type of summary of a frequency distribution includes measures of *dispersion* such as the *standard deviation* and the *coefficient of variation.* These measures are used to compare the relative wideness of spread in any two or more frequency distributions. Their characteristics, applications, and methods of computation can be found in any elementary textbook on statistics and will not be considered here. They are mentioned merely as being widely used ways of summarizing frequency distributions.

Comparing frequencies. A common and simple method of comparing frequencies is the use of the *ratio.* A ratio is merely an indicated or actual quotient which relates the size of one number to another.[4] Their chief utility is to act as a relative measure and thus permit the comparison of otherwise unequal numbers. For example, if we wish to know the relation of the female to the male death rates at various ages, a series of ratios, as shown in the accompanying table, will help. This adds to the knowledge, shown by the second and third columns, that male death rates are generally higher than female, by showing that this female advantage is constant through life, except for being somewhat larger than usual at ages 5 to 9, and somewhat smaller at ages 75 and over. The figures in column

Deaths per Thousand Native White Population by Age and Sex in Cities of 100,000 or more, Ohio, 1930 *

Age	Male	Female	Ratio of Male to Female
0–4	17.2	13.6	1.26
5–9	2.4	1.7	1.41
10–14	1.5	1.2	1.25
15–19	2.4	1.9	1.26
45–54	12.3	10.0	1.23
55–64	25.8	19.9	1.30
65–74	59.1	45.9	1.29
75 and over	135.6	114.5	1.18

* Adapted from Warren S. Thompson, *Population Problems* (New York: McGraw-Hill, 1942), p. 228.

---

[4] See Margaret Jarman Hagood, *Statistics for Sociologists* (New York: Holt, 1941), pp. 113–115.

four are actual quotients; they could have been expressed as 172 to 136, 24 to 17, 15 to 12, etc., but it is the actual quotients which are most useful since they reduce the right-hand figure to 1 in every case, thus allowing the comparison to be made easily.

A related method of comparing values is the *proportion*. This measure is a fraction such that the numerator is one of two observed frequencies and the denominator the sum of observed frequencies. Thus the previous table could have been expressed as the proportions of all deaths at various ages which were male deaths. Thus the ratios 1.26, 1.41, 1.25, etc., expressed as proportions would read 0.558, 0.585, 0.555, etc. When proportions are expressed in multiples of 100, they are *percentages*. Thus in the above example, 55.8 per cent of all deaths in large Ohio cities in 1930 between ages 0 and 4 were experienced by males, etc. Whether ratios, proportions, or percentages are used in the analysis of data is purely a matter of preference and of the way in which the researcher wishes to communicate his findings.[5]

The purpose of using percentages (and ratios and proportions as well) is to simplify the problem of comparison. It is important, therefore, to see exactly what their use implies so that they will not be misused. First, they can serve to put *qualitative* characteristics into numerical form. Thus it is possible to compare two college classes on the basis of sex by saying that one class is 60 per cent male and the other only 40 per cent. This is a perfectly good comparison, but care must be taken not to translate this statement by saying that one is more *masculine* than the other and thus implying that masculinity is a quantitative characteristic. Such errors in logic are common and easily fallen into, but the consequences may be disastrous.

Second, percentages reduce two frequency distributions to a common base, as was illustrated above, thus making comparisons much simpler. It must be noted, however, that this obscures *absolute* comparisons, and sometimes it is these rather than the relative comparisons which are important. Zeisel gives an example of this problem [6] as shown in the accompanying table. Here the absolute and the relative figures produce ap-

Comparative Growth of Cities *M* and *N*, 1941–1946

City	1941	1946	Increase	Percentage Increase
M	1,000,000	1,200,000	200,000	20.0
N	500,000	650,000	150,000	30.0

---

[5] See *ibid.*, pp. 115–123.

[6] Zeisel, *op. cit.*, p. 77.

parently opposite results. The problem, then, is which is "correct." Did city *M* grow more or less rapidly than city *N*? It is clear that this is a problem in logic. Thus it could be reasoned that, if the growth were "normal," we would expect the larger city to attract more people than the smaller and hence would say the percentages are the "significant" figure. If, on the other hand, the growth of cities depended not upon previous size but other factors such as the addition of new industries, then we might say that the larger city grew faster in the ratio of 200,000 to 150,000, or one and one-third times as fast.

In other words, while percentages are a great aid to communication through simplification, they are also susceptible to misuse by obscuring significant facts. This underlines the importance of never presenting a table of percentages without also showing the actual numbers which they represent. This not only guards against the kind of misinterpretation of the kind described above, but also assures the reader that there is an adequate base for the calculation of a percentage. For example, the statement that three of a total of only four people interviewed favored the Democratic party becomes very misleading in the form "three out of every four" or "75 per cent" of those interviewed favored the Democratic platform.

Bases for computing percentages. When a simple frequency distribution is turned into percentages it is usually for the purpose of making clear the comparisons between the several class intervals. In this case there is seldom a problem as to what the base of the percentages should be. One case, however, in which a difficulty may arise, is where the tabulation includes residual categories, or more than one dimension. An example will show this more clearly. If a sample polled in relation to a bond issue gave the results shown in the accompanying table, there would be a problem of interpretation. Thus if the base for percentaging were taken as the

Responses to Question, "Do You Favor the School Bond Issue Which Will Be Voted On in November?"

Responses	*Number*
Yes	97
No	78
Undecided	44
Refuse to Answer	31
Total	250

total sample, those favoring would be reported as 39 per cent; if the base were all *those replying,* the result would be 44 per cent; and if it were taken as *those who would answer and who had made up their minds,* it would be 56 per cent. It is very important, therefore, that the base of a percentage be fully and carefully described to avoid confusion.

The more common problem arises when a cross tabulation is used, for in this case, even though the categories are complete, it is still necessary to make a choice as to whether the percentages should be computed horizontally or vertically. Zeisel gives a general rule which can usually be followed, by stating[7] that *"the percentages should be computed in the direction of the causal factor."* This does not mean that one of the factors *must be* the actual cause of the other, but merely that in the mind of the analyst one of them is thought of as influencing the other. The illustration given by Zeisel makes this clear. First of all, percentages may be calculated to either the vertical or the horizontal totals. Thus the figures given in the first of the accompanying tables could be expressed in percentages calculated vertically, so as to appear as shown in the second table. Such a presentation shows the racial composition of two groups, those dying of cancer and those dying of some other cause. This is a clumsy way of presenting the analysis, since we are not likely to think the cause of death will affect materially the relative number of whites and Negroes. Rather it is more probable that the hypothesis was that for various reasons death accompanied by a diagnosis of cancer is more common for one group than another. Therefore the percentages should be computed horizontally, as shown in the third table. Such an arrangement points directly to the de-

Deaths from Cancer in the United States by Race

RACE	CAUSE OF DEATH		
	Cancer	All Others	Total
White	139,627	1,055,804	1,195,431
Negro	9,182	169,391	178,573
Total	148,809	1,225,195	1,374,004

Race	Cancer	All Others	Total
White	93.8	86.2	87.0
Negro	6.2	13.8	13.0
Total	100.0	100.0	100.0

Race	Cancer	All Others	Total
White	11.7	88.3	100.0
Negro	5.1	94.9	100.0
Total	10.7	89.3	100.0

---

[7] *Ibid.*, p. 88.

sired Negro-white comparison, rather than requiring still further analysis as did the other mode of calculating the percentages.

It will frequently be the case that this "causal principle" will not be applicable as clearly as in the above case. For example, suppose the number of years of school completed were being cross-tabulated with "economic worth" for a series of adults; would higher education be considered as the "cause" of higher economic standing, or vice versa? Clearly it could be viewed either way. The general principle, then, is to compute in the direction of the factor which the analyst wishes to emphasize.

That the analyst's concept of causality or desire for emphasis is not always the sole determiner of the direction in which percentages should be computed, however, is shown by another principle stated by Zeisel.[8] This principle is that percentages should be run *only in the direction in which a sample is representative.* The illustration of this rule deals with the collection of political preferences from a representative sample of 8,000 persons in each of two states, with the results shown in the accompanying table. If these results are analyzed from the point of view that differences between the states are "causal" for differences in party allegiance, then the table would be percentaged horizontally, as shown in the second table. Such results would be "sensible" in showing that party *YY* was in a favorable position in both states, but especially so in state *A*.

STATE	FAVOR PARTY		Total
	XX	*YY*	
A	2,500	5,500	8,000
B	3,500	4,500	8,000

STATE	PERCENTAGE FAVORING PARTY		Total
	XX	*YY*	
A	31.0	69.0	100.0
B	44.0	56.0	100.0

If the original question were reversed, however, and it were hypothesized that the differences in the *parties* rather than the differences in the *states* was the "cause" of the distribution, the percentages would have been calculated vertically, as shown in the third table. If this table were an

---

[8] *Ibid.*, p. 101.

STATE	PERCENTAGE FAVORING PARTY	
	XX	*YY*
A	42.0	55.0
B	58.0	45.0
Total	100.0	100.0

accurate analysis it would seem to say that 42 per cent of the votes of party *XX* would come from state *A*. If it is recalled, however, that the samples were representative of each state and *not* of the two parties, it is clear that these percentages are meaningless. Since we do not know the relative contribution of the two states to the total vote, no such percentages can be meaningful. Therefore where cross tabulations are representative in only one direction, percentages can be computed only in that same direction.

Presenting complex tables. Several examples have been given in which two-dimensional tables or simple cross tabulations have appeared. Often such tables are not adequate for the problem at hand where more than two dimensions must be simultaneously presented. The example of the size of houses given at the beginning of the chapter is a case in point. There it was pointed out that a complete presentation would involve at least six tables placed side by side. It is clear that such a presentation would be exceedingly confusing and thus poses the problem of simplifying the presentation. Suppose we had only three variables to deal with: size of family, size of house, and economic status. Suppose further that we are provided with the data shown in the accompanying table.

The Relation of Size of House to Size of Family, by Economic Status

INCOME	LARGE FAMILIES			MEDIUM FAMILIES			SMALL FAMILIES		
	Small House	Medium House	Large House	Small House	Medium House	Large House	Small House	Medium House	Large House
High	5	35	60	5	70	25	10	60	30
Medium	25	65	20	10	65	25	40	55	5
Low	35	55	10	35	55	10	75	25	0

One way of reducing the table is to remove one variable. Now any dichotomous percentage can be expressed by one figure only, since the second is then determined. If we say 65 per cent of a group are men it is not necessary to state also that 35 per cent are women. Thus if it is possible to turn any of the trichotomous variables into dichotomous ones, the table can be greatly simplified. So suppose the analyst's interest is in showing the problem of crowding, and he therefore rewords the table title and reenters the data as shown in the second table.

Percentage of Families of Various Sizes, by Income Groups, Residing in Small Houses

INCOME	FAMILY SIZE		
	Small	Medium	Large
High	10.0	10.0	5.0
Medium	40.0	10.0	25.0
Low	75.0	30.0	40.0

It would be possible to conclude from these data that income, more than family size, determined the proportion of small houses occupied.

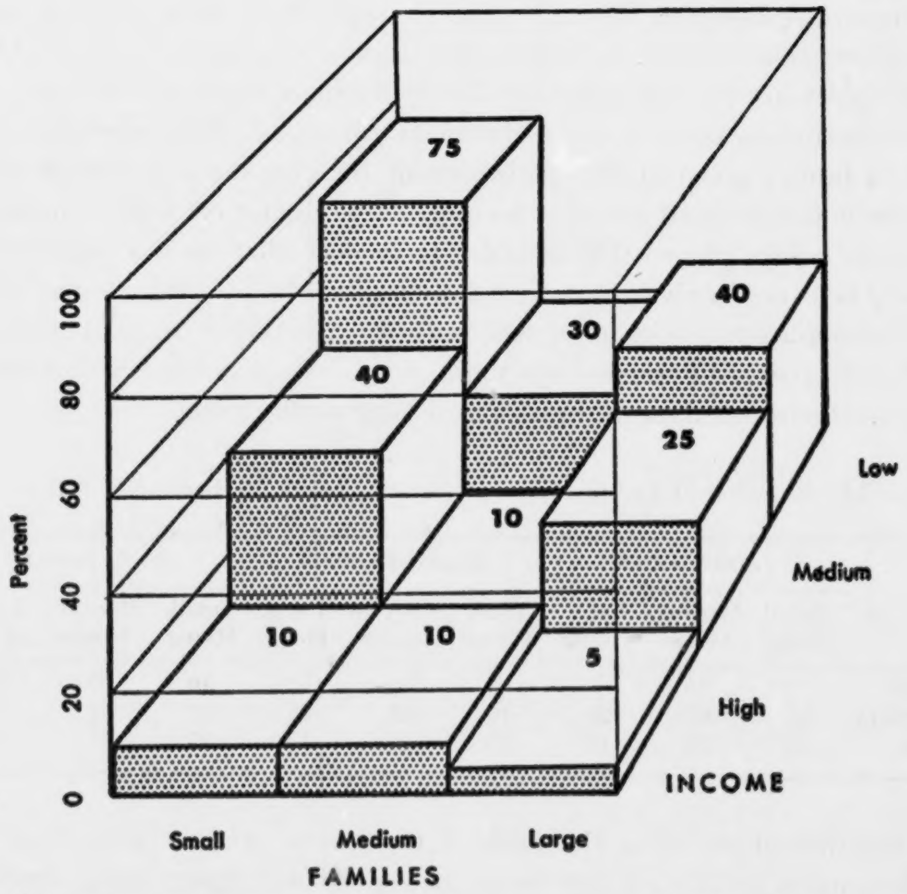


FIGURE 1. Graphic presentation of a trichotomous cross tabulation (income and size of families) for families living in small houses.

Figure 1 graphically illustrates what such a tabulation actually does to the data by showing the three-dimensional character of the cross tabulations. However, turning a trichotomy into a dichotomy is difficult and sometimes obscures the real point. Another method of summarizing the distribution, it will be recalled, is the use of an average. This table, if the values of "small," "medium," and "large" houses are known, can be simplified by the use of an average such as the mean, as shown in the third table.

Average Number of Rooms in Houses for Families of Various Sizes, by Income Groups

INCOME	FAMILY SIZE		
	Small	Medium	Large
High	4.7	4.6	5.7
Medium	3.2	4.5	4.4
Low	2.1	3.4	3.4

Again, the table becomes quite readable. The logic involved should be clear. In both cases the cell entries have been changed from frequencies or percentages, representing every category of the frequency distribution, into *single values,* each *standing for* the entire distribution.

General problems of presenting frequency distributions. This section has been concerned primarily with showing some of the ways of summarizing, simplifying, and presenting frequency distributions. Some of the more common problems have been spelled out and others indicated as belonging to the realm of statistics. With these in mind, the next section turns to some of the logical problems underlying simple analyses by cross tabulation.

LOGICAL ORDERING OF DATA

The cross tabulation of two or more attributes or variables is merely a formal and economical method of arranging the data so that the logical methods of proof may be applied. Thus, the methods of agreement, differences, or concomitant variation (correlation) may all be used in drawing conclusions from a cross tabulation. This should further underline the point made earlier in this chapter, that what can be gained through cross tabulation depends entirely on the logical design of the study and the insightfulness of the investigator. There are no "tricks" of cross tabulation which can guarantee that an analysis will provide the most significant and meaningful results possible.

It is possible, however, to discuss some of the modes of thought which lead to fruitful cross tabulation. First of all, the use of cross tabulation is, in effect, an approximation of the controlled experiment. This means that the analyst is really thinking in terms of cause and effect. That is, he has in mind one or more variables, variation in which can be used to *explain* variation in another variable. These "causal" dimensions are termed *independent variables* and the values to be explained are called *dependent variables*. It is because the analyst is really thinking in causal terms that there is a need for elaborating survey results through cross tabulation. If all that the analyst desires is an empirical statement of concomitant variation, simple cross tabulations will suffice. The need, however, to know the "meaning" of a relationship leads to asking many other questions which can be answered only by introducing other variables into the analysis.

Kendall and Lazarsfeld have given a classification of three types of elaboration which they call *interpretation, explanation,* and *specification.*[9]

Interpretation. This is the process of stratifying a previous cross tabulation by another variable (called a *test variable*) under circumstances such that the test variable has occurred (1) subsequent in time to the independent variable, but (2) prior to the dependent variable. One other requirement is that the test factor be correlated with both independent and dependent variables. Such a procedure allows the further interpretation of an original relationship in terms of a third factor, in this case called an *intervening* variable.

This may be diagrammed as follows:

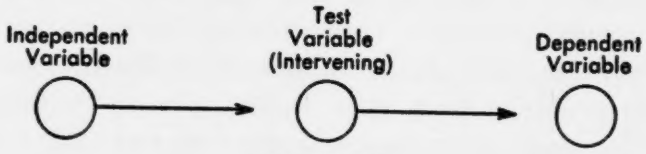


FIGURE 2

For example, when the men students at a private university were cross-tabulated by whether they attended public or private secondary schools and by their university academic records, those attending public schools showed a superior record. On the face of it, such a finding would suggest an explanation of the dependent variable (grades) in terms of the excellence of preparation in the independent variable (public versus private secondary schools). Anyone conversant with the admissions policy of private universities might suspect, however, that events could have occurred between secondary training and acceptance at the university which might give the finding an entirely different meaning.

Thus an intervening variable in terms of excellence of academic work in secondary school could logically be introduced. If there were circumstances in the application procedure which meant that public school graduates had to meet higher academic standards than private school graduates, the original correlation between the type of secondary school and subsequent grades would disappear. In this case, the test factor, high school record, would serve to interpret the original finding since when men of the *same high school achievement level* are compared, the relation of type of school to undergraduate success would disappear.

---

[9] Patricia L. Kendall and Paul F. Lazarsfeld, "Problems of Survey Analysis," *Continuities in Social Research,* Robert K. Merton and Paul F. Lazarsfeld, eds. (Glencoe, Ill.: Free Press, 1950). The entire paper is relevant here, but especially pp. 148–158.

Explanation. This is essentially the same kind of reasoning as interpretation in that it seeks to reduce an originally observed correlation through the use of a test factor. In this case, however, the test factor is an *antecedent variable* rather than an intervening variable. In other words, an observed relationship is explained in terms of a third factor which occurred earlier than the independent variable and which is related both to it and the dependent variable.

This may be diagrammed as follows:

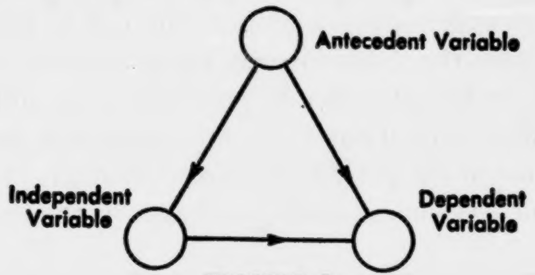


FIGURE 3

For example, a study of birth rates in Puerto Rico produced a cross tabulation which showed a negative correlation between economic status and the birth rate. Both the variables in this case were measures of adult life experience, and since "causal" reasoning was involved, the question was asked as to whether or not there was some variable characterizing the earlier portion of the life history which might "explain" this correlation. Consequently education was used as an antecedent test variable, with the result that the correlation between economic status and birth rates all but disappeared when cross-tabulated separately for comparable educational levels. In this case, then, the use of a test factor "explained away" an original observation and substituted another for it. We should describe the relationship between economic status and birth rates in Puerto Rico as a *spurious* correlation. This raises the question, "How can we be certain that the relationship between education and birth rates is actually a causal one and not merely another spurious correlation?" The only answer to this is to say that if the application of all known relevant test factors fails to reduce this correlation materially, the assumption of causality can usually be made.

Specification. Whereas in explanation and interpretation the question asked is whether or not the use of a test variable will cause the sharp reduction or disappearance of an observed relationship, the goal of specification is quite different. In this case the "causal" sequence is modified by specifying varying conditions under which the original correlation will exist in greater or lesser intensity. The example given by Kendall and

Lazarsfeld uses the findings from *The American Soldier*[10] shown in the accompanying table.

Educational Level

Rank	High School Graduate or Better	Less than High School Graduate
Non-com.	61.0	43.0
Pvt., Pfc.	39.0	57.0
Total cases	3,222	3,152

A crude measure of the relationship between education and rank is taken to be the difference between 61 per cent and 43 per cent. This value, called *f*, is 0.18. Now the question was asked whether an *f* value of 0.18 would be likely to be found under all circumstances, and it seemed logical that the answer was "no." That is, opportunities for promotion may not have been the same at all periods of World War II. Hence a cross tabulation by length of service was made, as shown in the second table.

Rank	Have Served for Less than Two Years		Have Served for Two Years or More	
	High School Graduate or Better	Less than High School Graduate	High School Graduate or Better	Less than High School Graduate
Non-com.	23.0	17.0	74.0	53.0
Pvt., Pfc.	77.0	83.0	26.0	47.0
Total cases	842	823	2,380	2,329
	$f = 0.06$		$f = 0.21$	

Since the *f*'s are quite different for the two length-of-service categories, the analyst can now state that the relation between education and rank is greater for those who entered the Army early than for those who entered it later.

Specification, then, is the process of describing the conditions under which a particular relationship may exist or not exist, or may exist to a greater or lesser degree. Like explanation and interpretation, therefore, it makes any "causal" analyses more acceptable and meaningful.

General aspects of elaboration. There seem to be two major problems with respect to this aspect of analysis. The applications of test factors in the three preceding paragraphs doubtless seem simple enough. The two practical problems, however, are (1) whether or not to elaborate, and (2) what test variables to apply.

---

[10] *Ibid.*, pp. 163–164.

With respect to the first problem it can be pointed out that unexpected consequences may impel the analyst to elaborate. If the findings conform to anticipation, there is little stimulus for further analysis. This is a somewhat dangerous situation, for it may result in the acceptance of spurious or uninterpreted relationships. As Kendall and Lazarsfeld put it:

> "Our first concern is whether or not the relationship is a spurious one; consequently our initial efforts of elaboration are usually of the explanatory type. Once we have gained some assurance that the original relationship is not a spurious one, we try to interpret the result or to specify it.[11] We ask ourselves what variables might provide the links between the 'cause' and the 'effect,' or what conditions might show the original relationship to be even more pronounced than we originally saw it to be. The elaboration of a particular result can go on almost indefinitely. We are limited only by our lack of ingenuity in thinking of factors by which to elaborate the result, by the absence of data to check the relevance of factors which we have thought of, or by the difficulties of dealing with few cases as the process of elaboration is extended."

Elaboration, therefore, is a process which is limited only by the analyst, on the one hand: his ability, his patience, and his purposes; and, on the other hand, by the nature of his data.

Unfortunately, it is not possible to give concise directions for selecting the factors to be used as test variables. It should be recalled that the selection of every test variable actually constitutes the formation of a hypothesis which can be tested by the data at hand. Therefore all the materials in the chapters on hypotheses and experimental design will apply here.

In other words, the ingenuity of the analyst coupled with his knowledge of the field in which he is working will be the major source of test variables. Two methods of inducing such hypotheses, however, may be mentioned. First, it would be possible in some studies to cross-tabulate routinely every factor with every other factor. In practice this is usually clumsy and the number can be cut down substantially by reference to existing theory. In any case, wide ranges of cross tabulations will often suggest the selection of test variables. Second, the study of deviant cases will also often indicate the kind of interpretation, explanation, or specification which will improve the analysis.

In the most general sense it must be said that the whole process of analysis is not so much a matter of manipulative techniques as it is of the rigorous application of the basic principles of scientific method. The research worker who is fully acquainted with the problems of *designing* research will have fewer troubles in *analyzing* his data.

---

[11] We must stress that the failure of one antecedent test factor to disqualify the causal connection between two variables is in no way a definitive demonstration of that causal relationship.

CHAPTER SEVEN

CLASS NOTES

SOME ILLUSTRATIVE EXERCISES

1. Random Selection of Experimental and Control Groups

A research project is being designed to test the effectiveness of a program for the dissemination of community information by means of a newsletter. One group of 20 households (the experimental group) will receive the newsletter weekly for a period of six months. Another group of 20 households (the control group) will not receive the newsletter but will be left exposed to the normal flow of community information.

Select the experimental and control groups from the roster and /or map of community households provided in the class notes in Chapter Four. Show the research design in tabular form.

2. The Judgment of Representativeness

Are the groups you have selected representative samples? Indicate the basis for your answer:

Are the groups you have selected representative samples? Indicate the basis for your answer:

(a) Assume that a summary information roster is available. (See the end of the roster.)

(b) Assume that there is no summary information.

3. The Validity of Data Collection

A basic variable in the research project is socio-economic status.

Operationally define five different ways to determine the socio-economic status of the households.

4. The Assumption of Rectilinearity

One of the hypotheses to be tested involves the relationship between age of the head of the household and the number of hours of television watched per week. The members of the research team are uncertain, however, whether the relationship would be rectilinear or curvilinear. Inspect the following data (constructing a scatter diagram if necessary).

TABLE I

AGE OF HEAD OF HOUSEHOLDS AND NUMBER OF HOURS OF TELEVISION WATCHED PER WEEK

AGE	NUMBER OF HOURS
23.9	1.4
21.3	2.2
31.1	2.6
25.2	1.9
22.5	2.0
31.8	3.4
25.3	.6
31.5	3.3
22.9	1.5
23.5	1.25
30.5	2.1
31.2	3.0
24.2	.8
27.6	1.0
31.5	2.8
28.9	1.5
28.5	.9
30.6	2.7
29.9	1.7
29.8	2.2

(a) Is the assumption of rectilinearity justified?

(b) Is the correlation of age and hours of watching television: High; Low; Non-Existent?

5. Variables Into Attributes

For the purposes of analysis, we wish to simplify the data given in #4 above. Convert the variables of age and hours of watching television into dichotomous attributes. Justify your cutting points.

6. Building Typologies

In a study of 2,200 individuals characterised by psychiatrists as being likely to attempt suicide, it was found that there were 1,100 individuals who had experienced great emotional stress in the recent past, and 53.6% of those had tried to commit suicide. Of those not experiencing great emotional stress in the recent past, 46.4% had tried to commit suicide.

It was also found that of those individuals described as suffering from social isolation--there were 1,100 of those too--as it turned out, 590 tried to commit suicide. Of those individuals not suffering from social isolation, 510 tried to commit suicide.

Among those individuals who had experienced emotional stress, there were 1,000 who had not suffered from social isolation and 500 of these had attempted suicide. Of all those individuals who had suffered from social isolation in the group studied, there were 1,000 who had not experienced emotional stress and 500 of these attempted suicide.

The research team concluded that there was very little relationship between attempted suicide and emotional stress or social isolation, when stress or isolation were present by themselves. When stress and isolation were both present, however, the probability of attempted suicide increased sharply.

Present the data bearing on this hypothesis in tabular form.

7. Measuring a Relationship

(a) Variables

A study of the causes of delinquency has shown that both age and income of parents are related to the number of serious delinquent acts committed by the individual. The question to be settled at this point is which of these two variables (age or income) is most closely correlated with delinquency.

Make your decision on the basis of the following data:

(1) The total variation in the number of delinquent acts equals 20.

(2) The total of the variation in the number of delinquent acts explained on the basis of income equals 16.

(3) The total of the variation in the number of delinquent acts explained on the basis of age is 14.

Is the larger correlation coefficient (r):

Very low or negligible	(.00 ± .20)
Low, present but slight	(± .20 to ± .40)
Substantial or marked	(± .40 to ± .70)
High or very high	(± .70 to ± 1.00)

(b) Attributes

Using Yule's Q ($Q = \frac{ad - bc}{ad + bc}$), Measure the association in the following two tables:

Which table shows the higher correlation?

(Note: Show your calculations)

TABLE I

	INCOME	
	High	Low
Age 25 - 45	20	10
46 or more	10	20

TABLE 2

		INCOME	
		High	Low
Race	Negro	15	30
	White	30	15

8. Testing for Significance of Difference

A community project to reduce unemployment is being based in part on a study of the job histories of workers at a number of large manufacturing companies in the area. The data show that the mean age of male workers who have been fired in the recent past is 31.97; the mean age of female workers fired is 33.39.

The data are based on a random sample of 713 unemployed men and a random sample of 287 unemployed women. The immediate question is whether the difference of 1.42 years in the mean age of male and female workers who have been fired is simply due to sampling vagaries or represents a real difference in the populations of male and female workers which might have to be taken into account in the community project.

(a) State the null hypothesis in full.

(b) Test your hypothesis in light of the following data:

Standard Error of the Difference
Between Means = .370

95% of the normal distribution will fall within 1.96 standard deviations above and below the mean.

99% of the normal distribution will fall within 2.58 standard deviations above and below the mean.

(c) State your conclusion about your hypothesis, with specific reference to the level of significance.

9. The Influence of an Intervening Characteristic

A survey of lawyers in private practice reveals that the longer the lawyer has been in practice the higher his annual income. (The data are given below.)

ANNUAL INCOME

Length of Practice	$15,000 per Year or Less	More than $15,000 per Year
0 - 10 Years	700	300
11 Years or More	510	690

It is hypothesized that a major reason for this association between length of practice and income is that the longer a lawyer is in practice, the more likely he is to become a partner in a firm--and such partners will tend to have higher incomes than non-partners.

The necessary information is given below:

Length of Practice	Annual Income	Partner	Total Number of Lawyers
11 Years or More	More than $15,000	Yes	640
11 Years or More	More than $15,000	No	50
11 Years or More	$15,000 or Less	Yes	160
11 Years or More	$15,000 or Less	No	350
0 - 10 Years	More than $15,000	Yes	200
0 - 10 Years	More than $15,000	No	100
0 - 10 Years	$15,000 or Less	Yes	100
0 - 10 Years	$15,000 or Less	No	600

10. The Influence of an Antecedent Characteristic

A team of participant observers, working with a government agency, have come to the conclusion that executives who are best able to handle criticism from those who supervise them also tend to be executives who get along best with their subordinates.

(The data age given below.)

	Ratings of Executives on Responses to Criticism		
Ratings of Executives on Relations with Subordinates	Good	Poor	Total
Good	50	40	90
Poor	40	50	90

It is hypothesized that the observed relationship between the two ratings is due in part to a management training program which was provided for one-half of the group. That is to say, the "good" ratings on both characteristics are caused by the management training, but outside of that there is no inherent connection between them.

The information necessary to test this hypothesis is given below:

Management Training Program	Relations with Subordinates	Reactions to Criticism	
Yes	Good	Good	40
Yes	Good	Poor	20
Yes	Poor	Good	20
Yes	Poor	Poor	10
No	Good	Good	10
No	Good	Poor	20
No	Poor	Good	20
No	Poor	Poor	40

Indicate if the hypothesis is correct, using the necessary tables to substantiate your conclusion.

Is there any correlation between ratings, with exposure to the management training program held constant?

11. Specification of Conditions

I. A study has been made which indicates that families presently on welfare tend to be more permissive in child-raising than families which were on welfare two years ago. (The data are given below.)

	Permissive Child-Rearing Practices		
	Yes	No	Total
Families Now on Welfare	180	120	300
Families on Welfare 2 Years Ago	120	180	300

The data indicate, however, that families now on welfare differ from families on welfare two years ago in terms of the age of the head of the household; and since age of the head of the household may be related to child-rearing practices, it has been suggested that the age of the head of the household be held constant in making comparisons.

When age is held constant, is the proportion of families with permissive child-rearing practices among those families now on welfare greater than, less than, or equal to the proportion among those families on welfare two years ago?

The necessary information is given below:

	Permissive Child-Rearing Practices		
	Yes	No	Total
Families Now on Welfare			
Age of Head			
20-24	72	18	90
25-29	88	22	110
30-34	12	48	60
35-40	8	32	40
Families on Welfare 2 Years Ago			
Age of Head			
20-24	36	9	45
25-29	44	11	55
30-34	22	88	110
35-40	18	72	90

If families now on welfare had the same age distribution as families on welfare two years ago, what proportion of families now on welfare would have permissive child-rearing practices?

I I. In examining the relationship between race and perception of occupational opportunities, it is found that there is no difference between the two groups in the community studied. (The data are given below.)

Race	Perception of Occupational Opportunities		
	Many Opportunities	No Opportunities	Total
White	160	340	500
Negro	160	340	500

The data indicate, however, that the two random samples of whites and Negroes differ with respect to individuals' belief in the availability of blue collar jobs. Since belief in the availability of blue collar jobs may influence the individual's perception of occupational opportunity, it has been suggested that the belief in availability of blue collar jobs be held constant in making comparisons.

When belief in the availability of blue collar jobs is held constant, do Negroes and whites differ with regard to their perception of occupational opportunity?

The necessary information is given below:

	Perception of Occupational Opportunities		
	Many Opportunities	No Opportunities	Total
Negroes			
Blue Collar Jobs Available	80	20	100
Blue Collar Jobs Not Available	80	320	400
Whites			
Blue Collar Jobs Available	80	320	400
Blue Collar Jobs Not Available	80	20	100

If Negroes were distributed in the same fashion as whites with respect to belief in the availability of blue collar jobs, what proportion of Negroes would believe that there were many occupational opportunities?

APPENDIX A

BIBLIOGRAPHY OF ADDITIONAL READINGS AND METHODOLOGY

Abramson, Norman. *Information Theory and Coding.* New York: McGraw-Hill, Inc., 1963.

Ackoff, Russell L. *The Design of Social Research.* Chicago: University of Chicago Press, 1953.

Alluisi, Earl A. *Basic Fortran for Statistical Analysis.* Homewood: The Dorsey Press, 1967.

The American Journal of Sociology. September, 1956. Vol. LXII, #2. Special Issue on Interviewing.

Anderson, R. L. "Distribution of the Serial Correlation Coefficient," *Annals of Mathematical Statistics*, 13, 1942, pp. 1-13.

Arley, Niels and Buch, K. Rander. *Introduction to the Theory of Probability and Statistics.* New York: John Wiley and Sons, 1950.

Backstrom, Charles H. and Hursh, Gerald D. *Survey Research.* Evanston: Northwestern University Press, 1963.

Barclay, George. *Techniques of Population Analysis.* New York: John Wiley and Sons, 1958.

Bereleson, Bernard and Stiener, Gary A. *Human Behavior: An Inventory of Scientific Findings.* New York: Harcourt, Brace and World, Inc., 1964.

Blalock, Herbert M. and Blaylock, Anne B. (ed.). *Methodology in Social Research.* New York: McGraw-Hill, Inc., 1968.

Blalock, Herbert M. *Social Statistics.* New York: McGraw-Hill, Inc., 1960.

Blalock, Herbert M. *Causal Inferences in Nonexperimental Research.* Chapel Hill, N. C.: University of North Carolina Press, 1964.

Blumer, Herbert. "Sociological Analysis and the Variable." American Sociological Review, Vol. 21, #6 (Dec. 1956). pp. 683-690.

Bonjean, Charles M., Hill, Richard J. and McLemore, S. Dale. Sociological Measurement and Inventory of Scales and Indices. San Francisco: Chandler Publishing Company, 1967.

Brogden, H.E. "Variation in Test Validity with Variation in the Distribution of Item Difficulties, Number of Items, and Degree of Their Intercorrelation," Psychometrika. 11, pp. 197-214, 1946.

Bruyn, Severyn T. The Human Perspective in Sociology. Englewood Cliffs: Prentice-Hall, 1966.

Campbell, Donald T. and Fiske, Donald. "Convergent and Discriminant Validation by the Miltitrait-Multimethod Matrix." Psychological Bulletin, Vol. 56, #2 (March 1959).

Campbell, Donald T. and Stanley, Julian C. Experimental and Quasi-Experimental Designs for Research. Chicago: Rand McNally and Company, 1966.

Campbell, Donald T. "Factors Relevant to the Validity of Experiments in Social Settings." Psychological Bulletin, Vol. 54, #4 (July, 1957). pp. 297-312. Bobbs Merrill Reprints #S-352.

Campbell, Donald T. "From Description to Experimentation: Interpreting Trends as Quasi-Experiments," in Chester W. Harris (ed.), Problems in Measuring Change. Madison, Wisconsin: University of Wisconsin Press, 1963.

Chernoff, Herman and Moses, Lincoln E. Elementary Decision Theory. New York: John Wiley and Sons, 1959.

Cicourel, Aaron V. Method and Measurement in Sociology. New York: Free Press, 1964.

Clark, Charles E. Random Numbers in Uniform and Normal Distribution. San Francisco: Chandler Publishing Company, 1966.

Coale, Ansley J. "Estimates of Various Demographic Measures Through the Quasi-Stable Age Distribution," in Millbank Memorial Fund, Emerging Techniques in Population Research. New York: pp. 175-193, May, 1963.

Coale, Ansley J. and Demeny, Paul. Regional Model Life Tables and State Populations. Princeton: Princeton University Press, 1966.

Cochran, W.G. "Analysis of Covariance: Its Nature and Uses." *Biometrics* 13, pp. 261-281, 1957.

Cochran, W.G. "Some Methods for Strengthening Common Methods." *Biometrics*, Vol. 10, No. 4, pp. 417-451, 1954.

Cohen, Morris and Nagel, Ernest. *An Introduction to Logic and Scientific Method*. New York: Harcourt, Brace and World, 1934. Chapters 2-4; 9-12.

Cohen, Arthur R. *Attitude Change and Social Influence*. New York: Basic Books, Inc., 1964. See Especially Chapter 9.

Coleman, James S. *Equality of Educational Opportunity*. Washington, D.C.: U. S. Government Printing Office, 1966.

Coleman, James. *Introduction to Mathematical Sociology*. New York: Free Press, 1964.

Courts, Frederick A. *Psychological Statistics: An Introduction*. Homewood: The Doresey Press, 1966.

Cronbach, L.J. "Coefficient Alpha and the Internal Structure of Tests." *Psychometrika*, Vol. 16, pp. 297-334.

Cronbach, Lee J. and Muhl, Paul E. "Construct Validity in Psychological Tests." *Psychological Bulletin*, Vol. 52 (May, 1955).

Draper, N.R. and Smith, H. *Applied Regression Analysis*. New York: John Wiley and Sons, 1966.

Doby, John T. (ed.). *An Introduction to Social Research*. (2nd ed.) New York: Appleton-Century-Crofts, 1967.

Edwards, Allen. *Experimental Design in Psychological Research*. (3rd ed.). New York: Holt, Rinehart and Winston, 1968.

Edwards, Allen L. *Techniques of Attitude Scale Construction*. New York: Appleton-Century-Crofts, 1957.

Entwisle, Doris. "Interactive Effects of Pretesting." *Educational and Psychological Measurement*, *21*. pp. 607-620, 1961.

Feldt, L.S. and Mahmoud, Moharram W. "Power Function Charts for Specification of Sample Size in Analyses of Variance." *Psychometrika*, 23. No. 3, pp. 201-210, 1958.

Feller, William. *An Introduction to Probability Theory and Its Applications.* 2nd Ed. New York: John Wiley and Sons, Inc., 1957.

Festinger, Leon and Katz, Daniel (eds.). *Research Methods in the Behavioral Sciences.* New York: Holt, Rinehart, and Winston, 1953.

Fisher, Ronald A. *Statistical Methods for Research Workers.* Edinburgh: Oliver and Boyd, 1958.

Francis, Roy G. "The Nature of Scientific Research," in John T. Doby (ed.), *An Introduction to Social Research* (2nd ed.). New York: Appleton-Century-Crofts, 1967. pp. 3-21.

Freeman, Linton C. *Elementary Applied Statistics: for Students in Behavioral Science.* New York: John Wiley and Sons, Inc., 1965.

Fruchter, Benjamin. *Introduction to Factor Analysis.* Princeton: D. Van Nostrand Company, 1954.

Garner, W.R. *Uncertainty and Structure as Psychological Concepts.* New York: John Wiley and Sons, Inc., 1962.

Goode, William J. and Hatt, Paul K. *Methods in Social Research.* New York: McGraw-Hill, 1952.

Goodman, L.A. "On the Multivariate Analysis of 3 Dichotomous Variables." *American Journal of Sociology.* 71, pp. 290-301. 1965.

Goodman, L.A. and Kruskal, William H. "Measures of Association for Cross Classifications." *American Statistical Association Journal*, 49, No. 268, pp. 732-764, 1954.

Goodman, L.A. and Kruskal, William H. "Measures of Association for Cross Classifications. Further Discussion and References." *American Statistical Association Journal,* 54. No. 2[illegible]5, pp. 123-163, 1959.

Gorden, Raymond L. *Interviewing: Strategy, Techniques, and Tactics.* Homewood: The Dorsey Press, 1969.

Gordon, R.A. "Issues in Multiple Regression." *American Journal of Sociology*, 73. pp. 592-616, 1968.

Gorow, Frank F. *Statistical Measures: A Programmed Text.* San Francisco: Chandler Publishing Company, 1962.

Green, Bert F. and Tukey, John W. "Complex Analyses of Variance: General Problems." Psychometrika 25. pp. 127-151, 1960.

Greenwood, Ernest. Experimental Sociology: A Study in Method. New York: King's Crown Press, 1945.

Greer, Scott. The Logic of Social Inquiry. Chicago: Aldine Publishing Company, 1969.

Guilford, J. P. Psychometric Methods. New York: McGraw-Hill, 1954.

Hagood, Margaret Jarman. Statistics for Sociologists. New York: Henry Holt & Company, 1952.

Hammond, Phillip E. (ed.). Sociologists at Work: Essays on the Craft of Social Research. New York: Basic Books, Inc., 1964.

Hauser, Philip M. and Duncan, Otis Dudley (eds.). The Study of Population: An Inventory and Appraisal. Chicago: University of Chicago Press, 1959.

Hayes, Margaret L. and Conklin, Mary E. "Intergroup Attitudes and Experimental Change." Journal of Experimental Education, pp. 26-27, 1953.

Hays, William L. Statistics for Psychologists. New York: Holt, Rinehart and Winston, 1963.

Hirschi, Travis and Hanan C. Selvin. Delinquency Research. New York: Free Press, 1967. Cf. Chapter 12, Reliability and Scaling."

Hovland, Carl I. "Reconciling Conflicting Results Derived from Experimental and Survey Studies of Attitude Change." The American Psychologist, 14. No. 1, pp. 8-17, 1959.

Huff, Darrell. How to Lie with Statistics. New York: W. W. Norton, 1954.

Hyman, Herbert, et al. Interviewing in Social Research. Chicago: University of Chicago Press, 1954.

Hyman, H. H. and Sheatsley, P. B. "Some Reasons Why Information Fails." The Public Opinion Quarterly II. pp. 413-423, 1947.

Hyman, Herbert. Survey Design and Analysis. New York: Free Press, 1955.

Hyman, Herbert. *Survey Design and Analysis: Principles, Cases, and Procedures*. Glencoe: Free Press, 1955.

Kaplan, Abraham. *The Conduct of Inquiry*. San Francisco: Chandler Publishing Company, 1964.

Katz, Leo and Proctor, Charles H. "The Concept of Configuration of Interpersonal Relations in a Group as a Time Dependent Stochastic Process." *Psychometrika 24*. No. 4, pp. 317-327, 1959.

Kerlinger, Fred N. *Foundations of Behavioral Research*. New York: Holt, Rinehart, and Winston, Inc., 1964.

Kish, Leslie. "Some Statistical Problems in Research Design." *American Sociological Review*, Vol. 24 (1959). pp. 328-338.

Kish, Leslie. *Survey Sampling*. New York: John Wiley and Sons, Inc., 1965.

Kitagawa, Evelyn M. "Standardized Comparisons in Population Research." *Demography*. Vol. 1, No. 1, pp. 236-315, 1964.

Knepp, Dennis L. "Table of $-p \log_2 p$ with Function and Argument to Four Decimal Places." Department of Social Relations, The Johns Hopkins University. 1967 (mimeo.).

Lazarsfeld, Paul F. and Rosenberg, Morris (eds.). *The Language of Social Research*. New York: Free Press, 1955. (Paperback)

Lindzey, Gardiner and Aronson, Elliot (eds.). *The Handbook of Social Psychology*. Reading: Addison-Wesley Publishing Company, 1969.
Vol. 1: Systematic Positions
Vol. 2: Research Methods
Vol. 3: The Individual in a Social Context
Vo. 4: Group Psychology and Phenomena of Interaction
Vol. 5: Applied Social Psychology

Madge, John. *The Origins of Scientific Sociology*. New York: The Free Press, 1962.

McGinnis, Robert. *Mathematical Foundations for Social Analysis.* Indianapolis: The Bobbs-Merrill Company, 1965.

McGinnis, Robert. "Randomization and Inference in Sociological Research." *American Sociological Review*, Vol. 22 (1957), pp. 408-416.

Merton, Robert K. *On Theoretical Sociology*. New York: The Free Press, 1967.
(Paperback) Includes:
"The Bearing of Sociological Theory on Empirical Research."
"The Bearing of Empirical Research on Sociological Theory."

Merton, Robert K. *Social Theory and Social Structure*. Glencoe: The Free Press, 1957.

Miller, Delbert C. *Handbook of Research Design and Social Measurement.* New York: David McKay, 1964.

Mood, Alexander M. and Graybill, Franklin A. *Introduction to the Theory of Statistics*. (2nd ed.) New York: McGraw-Hill, 1963.

Mueller, John H. and Schuessler, Karl F. *Statistical Reasoning in Sociology*. Boston: Houghton Mifflin Company, 1961.

Nagel, Ernest. *The Structure of Science: Problems in the Logic of Scientific Explanation*. New York: Harcourt, Brace and World, 1961.

Northrop, F. S. C. *The Logic of the Sciences and the Humanities*. New York: MacMillan, 1948.

Osgood, Charles E.; Suci, George J. and Tannenbaum, Percy H. *The Measurement of Meaning*. Urbana: University of Illinois Press, 1957.

Parten, M. *Surveys, Polls and Samples*. New York: Harper and Row, 1950.

Peatman, John G. *Introduction to Applied Statistics*. New York: Harper and Row, 1963.

Pelz, Donald C. and Andrews, Frank M. "Detecting Causal Priorities in Panel Study Data." *American Sociological Review, 29*. No. 6, pp. 836-848, 1964.

Phillips, Bernard S. *Social Research: Strategy and Tactics*. New York: MacMillan, 1966.

Ray, William S. *An Introduction to Experimental Design.* New York: MacMillan, 1960.

Richardson, Stephen A.; Dohrenwend, Barbara Snell and Klien, David. *Interviewing: Its Forms and Functions.* New York: Basic Books, Inc., 1965.

Riley, Matilda. *Social Research, A Case Approach*. New York: Harcourt, Brace and World, Inc., 1963.

Rose, Arnold M. *Theory and Method in the Social Sciences*. Minneapolis: University of Minnesota Press, 1954.

Rosenberg, Morris. *The Logic of Survey Analysis*. New York: Basic Books, Inc., 1968.

Rosenberg, Morris. *Society and the Adolescent Self-Image*. Princeton: Princeton University Press, 1965. pp. 33-36.

Rosenberg, Morris. "Test Factor Standardization as a Method of Interpretation." *Social Forces*, Vol. 41, #1 (1962). pp. 53-61.

Rozeboom, William W. *Foundations of the Theory of Prediction*. Homewood: The Dorsey Press, 1966.

Runkel, Philip J.; Smith, J. E. K. and Newcomb, Theodore. "Estimating Interaction Effects among Overlapping Pairs." *Psychological Bulletin* 54, No. 12, pp. 152-158, 1957.

Runyon, Richard P. and Haber, Audrey. *Fundamentals of Behavioral Statistics*. Reading: Addison-Wesley Publishing Company, 1967.

Scheffé, Henry. *The Analysis of Variance*. New York: John Wiley and Sons, Inc., 1959.

Sellin, Thorsten, and Wolfgang, Marvin E. *The Measurement of Delinquency*. New York: John Wiley and Sons, Inc., 1964.

Selltiz, Claire; Jahoda, Marie; Deutsch, Morton; and Cook, Stuart W. *Research Methods in Social Relations*. New York: Henry Holt and Company, 1959. (Revised One-Volume Edition).

Selvin, Hanan C. "A Critique of Tests of Significance in Survey Research." *American Sociological Review*, Vol. 23, #4 (Aug., 1958). Bobbs-Merrill Reprints #S-504. pp. 520-523.

Selvin, Hanan. "A Critique of Tests of Significance in Survey Research." *American Sociological Review*, Vol. 23 (1958). pp. 519-527.

Selvin, Hanan C. "Durkhiem's Suicide and Problems of Empirical Research." *American Journal of Sociology*, Vol. LXIII, #6 (May, 1958).

Shaw, Marvin E. and Wright, Jack M. *Scales for the Measurement of Attitudes*. New York: McGraw-Hill, 1967.

Sheps, Mindel C. "Applications of Probability Models to the Study of Patterns of Human Reproduction." Sheps, Mindel C. and Ridley, J. C. (eds.). *Public Health and Population Change: Current Research Issues.* Pittsburgh: University of Pittsburgh Press, pp. 307-332, 1965.

Siegel, Sidney. *Nonparametric Statistics for the Behavioral Sciences*. New York: McGraw-Hill, 1956.

Simon, Rita James (ed.). *The Sociology of Law: Interdisciplinary Readings*. San Francisco: Chandler Publishing Company, 1968. See Part IV: Contemporary Empirical Research.

Snedecor, George and Cochran, William. *Statistical Methods*, 6th ed. Ames, Iowa: Iowa State University Press, 1967.

Solomon, Richard L. "An Extension of Control Group Design." *Psychological Bulletin 46*. pp. 137-150, 1949.

Steele, Robert G. D. and Torrie, James H. *Principles and Procedures of Statistics*. New York: McGraw-Hill, 1960.

Stephan, F. F. and McCarthy, P. J. *Sampling Opinions*. New York: John Wiley and Sons, Inc., 1958.

Stevens, S.S. "Mathematics Measurement and Psychophysics." *Handbook of Experimental Psychology*. New York: John Wiley and Sons, Inc., 1951.

Stevens, S.S. "On the Theory of Scales of Measurement." *Science*. 103, #2684 (June 7, 1946). pp. 677-680.

Stinchcombe, Arthur L. *Constructing Social Theories*. New York: Harcourt, Brace, and World, Inc., 1968.

Stouffer, Samuel A., *et al.* *Measurement and Prediction*. Princeton: Princeton University Press, 1950.

Stouffer, Samuel A. *Social Research to Test Ideas.* New York: The Free Press, 1962.

Stouffer, Samuel A. "Some Observations on Study Design," in *American Journal of Sociology*, Vol. LV (Jan., 1950), pp. 355-361.
Available in *Bobbs-Merrill Reprints* #S-520. Also available in Miller, Delbert C. *Handbook of Research Design and Social Measurement*. New York: David McKay Co., Inc., 1964, pp. 20-26.

Tate, Merle W. and Clelland, Richard C. *Nonparametric and Shortcut Statistics*. Danville, Illinois: Interstate, 1957.

Tittle, Charles R. and Hill, Richard. "Attitude Measurement and Prediction of Behavior: An Evaluation of Conditions and Measurement Techniques." *Sociometry*, Summer, 1967, pp. 199-213.

Torgerson, Warren G. *Theory and Methods of Scaling*. New York: John Wiley and Sons, Inc., 1958.

Tukey, John W. "The Future of Data Analysis." *Annals of Mathematical Statistics 33*. No. 1, pp. 1-67, 1962.

United Nations. *Manual III, Methods for Population Projection by Sex and Age*. ST/SOA/Series A/25. New York: 1956.

United Nations. *Manual IV, Methods of Estimating Basic Demographic Measures from Incomplete Data*. ST/SOA/Series A/42. New York: October, 1967.

U.S. Bureau of the Census. *The Post-Enumeration Survey: 1950*. Bureau of the Census Technical Paper No. 4. Washington, D.C.: 1960.

Wallis, W. Allen and Roberts, Harry V. *Statistics: A New Approach*. New York: The Free Press, 1956.

Wallis, W. Allen. "Statistics of the Kinsey Report." *J.A.S.A. 44*. pp. 463-483, 1949.

Webb, Eugene J.; Campbell, Donald T.; Schwartz, Richard D.; and Sechrest, Lee. *Unobtrusive Measures: Nonreactive Research in the Social Sciences*. Chicago: Rand McNally and Company, 1966.

Westoff, Charles F.; Potter, Robert G., Jr.; and Sagi, Philip C. *The Third Child: A Study of Prediction in Fertility*. Princeton: Princeton University Press, 1963.

Whitla, Dean K. (ed.). *Handbook of Measurement and Assessment in Behavioral Sciences.* Reading: Addison-Wesley Publishing Company, 1968.

Wiener, Norbert. "A New Theory of Measurement: A Study in the Logic of Mathematics." *Proceedings of the London Mathematics Society, Ser. 2, 19.* pp. 181-205, 1920.

Winer, B. J. *Statistical Principles in Experimental Design.* New York: McGraw-Hill, 1962.

Young, Pauline V. *Scientific Social Surveys and Research.* (4th ed.). Englewood Cliffs: Prentice-Hall, Inc., 1966.

Zeisel, Hans. *Say It With Figures.* (4th ed.). New York: Harper and Row, 1947.

Zettersberg, Hans. *On Theory and Verification in Sociology.* (2nd ed.). Totowa, N. J." The Bedminister Press, 1963.